UNIVERSIDADE FEDERAL DO PARANÁ

RODRIGO MACIEL LOBO

O USO DE GRANDE VOLUME E VARIEDADE DE INFORMAÇÕES - *BIG DATA*

CURITIBA

2017

RODRIGO MACIEL LOBO

O USO DE GRANDE VOLUME E VARIEDADE DE INFORMAÇÕES - *BIG DATA*

Monografia apresentada como requisito parcial à conclusão do curso de graduação em Ciências Econômicas, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná.

Orientador: Prof. Adriana Sbicca

CURITIBA

2017

RESUMO

O grande volume de dados gerado por usuários tem se tornado um obstáculo, pois os paradigmas tradicionais de armazenamento e processamento não são mais adequados. Uma nova linha de modelos computacionais, algoritmos e componentes vem surgindo para atender esta necessidade.

Transações econômicas estão, atualmente, dependentes de alguma forma de computação que consiga capturar dados das transações para que sejam manipulados e analisados. A estatística convencional e técnicas econométricas, como a regressão, funcionam muito bem, e com a chegada do que denominamos como a era do *Big Data*, na qual há desafios de analisar grandes quantidades de dados de forma rápida, estas técnicas podem ter seu potencial amplificado a partir de novas ferramentas que permitem a leitura de grandes bases de dados.

A partir disto, é possível identificar como os conceitos dados ao *Big Data* permitem novas possibilidades, antes limitadas em função de peças de *hardware* para armazenamento de dados, abrangência da internet móvel, ferramentas analíticas para grandes conjuntos de dados, e como estes têm fornecido verdadeiros *insights* à chamada *Business Intelligence* (BI), uma economia de negócios ampliada a partir da gerência dos dados.

Neste trabalho serão apresentadas e exploradas estruturas que fundamentam o tema de *Big Data*. Serão ilustrados três estudos de caso utilizando-se da mineração de dados para a análise preditiva. O primeiro caso busca exemplificar uma forma de explorar os dados a partir de técnicas estatísticas, no caso a classificação, como forma de identificar possíveis novos clientes para adesão ao cartão fidelidade de uma empresa de comércio varejista realizando a análise de dados históricos com características de clientes que já aderiram. O segundo caso demonstra a utilização de dados históricos do Poduto Interno Bruto (PIB) do Brasil com o intuito de demonstrar tecnicamente o desenvolvimento de uma análise preditiva a partir do uso da regressão com o suporte de ferramentas para uso de *Big Data.* O terceiro caso busca, de forma hipotética, exemplificar uma das aplicações mais realizadas quando tratamos de *Big Data*, unir duas (ou mais) fontes de dados distintas com o intuito de prevenir o vírus Zika em

determinada região aliado à informação de nível de faixa de renda que o mesmo pode estar se proliferando. Isto será feito utilizando a captura de dados da internet em tempo real e dados do setor censitário disponibilizados pelo IBGE, alinhando estes dois conjuntos de dados para transformá-los em uma ação de análise preditiva.

Palavras-chave: *Big Data, Data Mining,* estatística, análise preditiva.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

BI	Business Intelligence
PIB	Produto Interno Bruto
SVM	Support Vector Machine
GLM	Generalized Linear Model
IBGE	Instituto Brasileiro de Geografia e Estatística
IPEA	Instituto de Pesquisa Econômica Aplicada

SUMÁRIO

INTRODUÇÃO

O presente trabalho tem como objetivo principal: explorar o conceito de *Big Data* e algumas técnicas e tecnologias que dão suporte ao uso do mesmo, como a mineração de dados, a estatística e a análise preditiva, e como este conjunto está alterando a forma de analisar informações devido ao grande volume e variedade de dados digitais sendo gerados a cada minuto. Esta nova forma de analisar envolve uma variedade de técnicas analíticas e estatísticas para desenvolver modelos que procuram predizer eventos ou comportamentos futuros.

Para se atingir o objetivo principal, busca-se atingir os seguintes objetivos específicos: i) explorar os conceitos de *Big Data*; ii) explorar os conceitos de Mineração de Dados; iii) realizar, por meio de um *software* de mineração de dados, testes para identificar o melhor modelo para prever um evento de aquisição de cartões fidelidade; iv) utilizar dados do Twitter em tempo real para identificar bairros específicos do Paraná onde há uma possível maior ocorrência do vírus Zika.

O capitulo um abordará o uso da internet pela sociedade e como a evolução de tecnologias existentes (smartphones, redes sociais) e o advento de novas tecnologias (Netflix, plataformas de *streaming* musical como Spotify) está impulsionando o uso da rede, desencadeando uma série de transformações econômicas e impulsionando mercados e países.

O capítulo dois explorará como esse montante de informações que está surgindo através dos dados vem sendo caracterizado, e como diferentes setores da indústria, como saúde, finanças, óleo e gás, entre outros, estão sendo impactados por tecnologias que agora trazem capacidade de análise de informações nunca antes percebidas. Também são exploradas as fases e as técnicas de análise em um ambiente governado por dados.

O capítulo três destacará uma das técnicas analisadas no capítulo anterior, a mineração de dados, e como funciona o processo de preparação dos dados para criação de um modelo.

No capítulo quatro serão aprofundados os principais modelos utilizados em mineração de dados, bem como seus algoritmos e exemplos de utilização no mundo real.

O capítulo cinco trabalhará com um exemplo de aplicação utilizando algoritmos e métodos estatísticos para analisar e prever o mercado de varejo e a fidelidade de seus clientes.

O capítulo seis irá demonstrar tecnicamente uma das técnicas mais aplicadas em *Big Data*, a regressão, com o suporte de uma linguagem de programação, a Linguagem R, com o objetivo de prever o PIB do Brasil nos próximos anos.

O capítulo sete, por fim, tem como objetivo demonstrar tecnicamente o uso de duas fontes de dados distintas para prever possíveis ocorrências de caso do vírus Zika com base em comentários feitos na rede social Twitter, utilizando técnicas de coleta de dados em tempo real e dados do setor censitário provenientes do IBGE.

1 A ERA DA COMUNICAÇÃO

A Internet nos permite alcançar um público sem precedentes e obter dados sobre ele de uma forma nunca antes imaginada. Castells (2000) descreve este momento vivido como sociedade em rede, a qual é vista como um emaranhado de informação circulante que tende a crescer através do mundo virtual. O avanço da tecnologia permite que a quantidade de informação produzida dia após dia aumente, ampliando-se a rede de conhecimento.

Com o surgimento da Web houve grande aumento no volume das informações eletrônicas, as quais trouxeram muitas vantagens quanto à possibilidade de troca, difusão e transferência de dados. Entretanto, este crescimento trouxe muitos problemas relacionados ao acesso, busca e recuperação das informações de real valor imerso em grandes volumes de dados (MODESTO, 2013).

A facilidade de aprendizado das novas tecnologias é boa parte responsável pela contribuição à crescente comunicação na era digital. Mesmo pessoas com menos conhecimento na área de tencologia conseguem aprender rapidamente a utilizar a rede, principalmente o que chamamos de redes sociais, como twitter, facebook e linkedin. Dados do IBGE demonstram como este crescimento foi expressivo especialmente nas classes mais baixas, onde o percentual de pessoas que acessam a internet passou de 3,8% em 2005 para 21,4% em 2011, e dados mais recente mostram que em 2014 o percentual de brasileiros com acesso à internet chegou a 54,8%.

Ribeiro (2014, p. 98), exemplifica o avanço e crescimento no volume de dados e informações que vêm se obtendo devido ao crescente uso de dispositivos móveis, de sensores industriais e biomédicos, fotos, vídeos, e-mails, redes sociais, comércio eletrônico, interações via *call centers*, dados públicos, informações sobre previsão de tempo, imagens médicas e outros dados científicos, câmeras para monitoramento, medidores inteligentes, GPS, aplicativos para troca de mensagens, aplicações que nos ajudam a pegar táxis, outras que nos ajudam na locomoção urbana evitando engarrafamentos, ou ainda no monitoramento de ônibus e aviões.

Segundo Setzer (1999), professor de ciência da computação do Instituto de Matemática e Estatística da USP, define-se *dado* “como uma sequência de símbolos quantificados ou quantificáveis. Portanto, um texto é um dado. Como são símbolos quantificáveis, dados podem ser armazenados em um computador e processados por ele. Também são dados: imagens, sons e animação, pois todos podem ser quantificados” Ainda segundo o autor, *informação* é *“*uma abstração informal que representa algo significativo através de textos, imagens, sons ou animação. Não é possível processar informação diretamente em um computador. Para isso é necessário reduzi-la a dados. A representação da informação pode eventualmente ser feita por meio de dados. Nesse caso, pode ser armazenada em um computador, mas o que é armazenado na máquina não é a informação, mas a sua representação em forma de dados.” Contudo, sabendo-se que informações são deixadas durante a navegação na rede e que são armazenadas pelas máquinas através dos dados, o conhecimento pode ser extraído destas informações. Melhorias na infraestrutura permitiram a migração dos dados de estruturas físicas para a nuvem, onde tudo pode ser acessado de onde estiver e quando quiser.

Porém, esses dados somente terão valor se puderem ser processados e analisados, produzindo conhecimento para seus detentores, podendo gerar impactos para qualquer segmento da economia e em escala. Em 2013, um relatório realizado por uma das maiores consultorias de negócios do mundo, McKinsey & Company, demonstrou que a área de ciência dos dados seria um dos principais catalisadores do crescimento econômico, chegando a compor 1,7% do PIB americano na previsão para 2020. *(*MCKINSEY*,* 2013).

Frost & Sullivan (2017) apontam que o mercado latino americano de *Big Data* responde atualmente por 5,1% do mercado global, movimentando US$ 2,48 bilhões na América Latina em 2016, com previsões de triplicar até 2022, atingindo a casa dos US$ 7 bilhões.

Neste cenário, impulsionado pelo grande volume de dados sendo gerados constantemente em grande velocidade e com a possibilidade de serem capturados de várias fontes geradoras aliando-se a possibilidade de interpretá-los, são os principais contribuintes que lançaram o *Big Data*.

Dados agora podem ser capturados em tempo real e transformados em *insights* de negócios, gerando valor aos indivíduos, empresas e governos. Fundos de Investimento usam Twitter para prever o desempenho do mercado de ações. Amazon e Netflix baseiam suas recomendações de produto das diversas interações de seus sites. Twitter, Facebook e Linkedin mapeiam o gráfico social das relações entre os usuários para entender mais sobre suas preferências e oferecer melhores produtos e serviços.

Empresas como Netflix têm utilizado a sua base de clientes que, a partir do uso de seu sistema, geram informações com potencial valor para criação de novos produtos a partir dos dados coletados. A série de televisão *House of Cards*, produzida em 2013 pela Netflix, foi realizada com base nas informações coletadas a partir de dados dos assinantes. (NY Times, 2013) Da mesma forma, a empresa Amazon utiliza o Kindle para criar seu sistema de recomendação e indicar livros de interesse para cada consumidor. (Investopedia, 2016).

2 O *BIG DATA*

Ainda que não haja uma definição específica para o conceito de *Big Data*, a ideia é a de que o volume de informação crescera tanto, especialmente com o advento das redes sociais, que as tecnologias atuais não conseguem processá-las. O termo corresponde à própria quantidade absurda de dados gerados atualmente. Anteriores à Era do *Big Data*, fórmulas matemáticas e técnicas de estatística e de probabilidade tinham capacidade reduzida de variáveis, pois eram realizadas por computadores tecnicamente inferiores para grandes cálculos em cima de um grande conjunto de dados. Com o desenvolvimento de processadores de alta capacidade e velocidade foi possível a criação de softwares mais poderosos que realizassem todos estes cálculos e transformassem esses dados em informações estratégicas para qualquer segmento da economia. Portanto, uma solução de *Big Data* trabalha com a agregação de algoritmos complexos de diversas fontes, de forma que se relacionem e gerem informações fundamentais que dão apoio à tomada de decisões das empresas e governos para o desenvolvimento de políticas públicas.

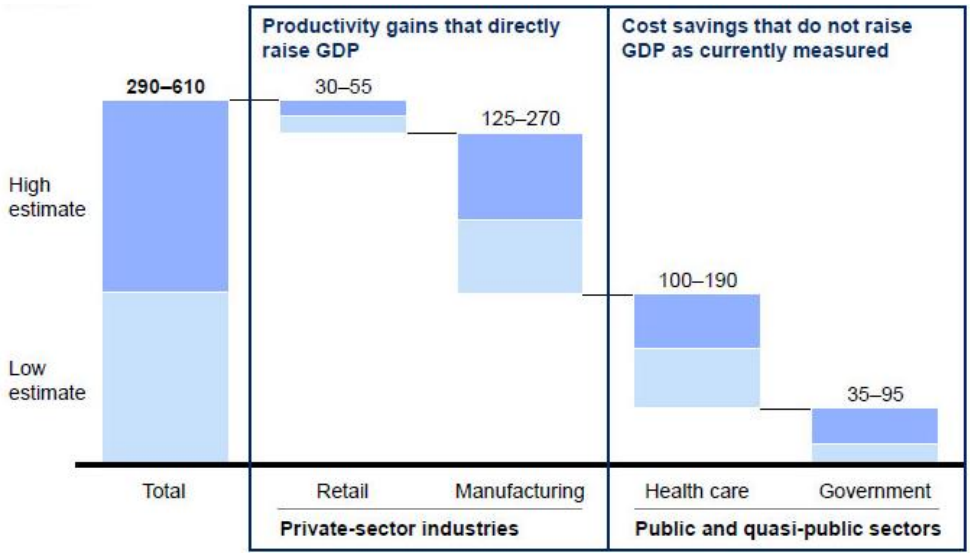
O início de suas aplicações remete-se à NASA no início da década de 1990, quando começou a utilizar *Big Data* para descrever imensos conjuntos de dados complexos incapazes de serem processados e analisados com os sistemas computacionais da época. Neste modelo, a captura, processamento e análise de dados eram feitos por meio de sistemas de alto impacto, envolvendo o trabalho simultâneo de diversas áreas da ciência. O objetivo destes sistemas era gerar conhecimento a partir dos dados brutos, que, quando analisados independentemente, não forneciam nenhum tipo de informação.

Segundo o professor Daniel Mendes da Data Science Academy, “*Big Data* é uma coleção de conjunto de dados, de grande volume e complexos, que não podem ser processados por bancos de dados ou aplicações de processamento tradicionais”. Muitos dados gerados possuem um tempo de vida curto, e se não forem analisados, perdem a utilidade. Dados são transformados em informação que precisam ser colocadas em contexto para que possam fazer sentido e gerar conhecimento. Muitos dados valiosos de diversos sistemas são descartados ou perdidos porque ninguém presta atenção a eles. Em outras palavras, *Big Data* é a

“capacidade de uma sociedade de obter informações de maneiras novas a fim de gerar ideias úteis e bens e serviços de valor significativo”. (MENDES, 2015).

Atualmente é muito comum a maioria das empresas utilizar o *Business Intelligence* (BI) como forma de analisar seus dados a partir da coleta, organização, criação de análises e compartilhamento com executivos. As informações geradas tornam-se fonte de apoio para a tomada de decisões. Contudo, o BI oferece informações qualitativas que permitem a tomada de decisão segura através de dados previamente trabalhados. Com a chegada do grande volume de dados, o BI não consegue trilhar o caminho por conta própria. Neste cenário, alia-se a mineração de dados, que utilizará algoritmos estatísticos para formar análises preditivas e em tempo real. Segundo Barbieri (2011, p.4) “A informática fez os dados e posteriormente transformou-os em informação. Agora o objetivo é usinar conhecimento a partir daquelas matérias-primas”. Desta forma, a inteligência de negócios e a mineração devem andar concomitantemente, tornando-se grande impulsionador econômico de vários segmentos.

Figura 1 - Ganhos de produtividade e redução de custos até 2020



Fonte: McKinsey Global Institute

A figura acima traz uma representação do impacto gerado no PIB dos Estados Unidos devido aos ganhos de produtividade gerados pelo *Big Data* até 2020, valor que chega até os $325 bilhões de dólares nos setores de manufatura e

varejo somados. Para o setor público, a previsão é de o *Big Data* contribuir com uma redução de custos de até $285 bilhões de dólares nos setores de saúde e governo.

Nas economias desenvolvidas da Europa, os administradores governamentais poderiam economizar mais de € 100 bilhões de euros (aproximadamente R$ 343 bilhões de reais) em melhorias de eficiência operacional usando *Big Data*, não incluindo o uso para reduzir fraudes e erros e aumentar a arrecadação de receitas tributárias. (McKinsey, 2011)

Diversos setores da indústria e políticas públicas se beneficiam de projetos de *Big Data*, alguns exemplo são: finanças, saúde, tecnologia, oléo e gás e infraestrutura, conforme serão exemplificadas algumas maneiras a seguir.

No setor de **finanças**, alguns exemplos de casos são a análise de dados através das linhas de negócios de uma instituição financeira, como empréstimo, seguro, produtos de cartões e serviços online para avaliação de mercado; análise de risco e predição de receita de novos produtos; análise de tendência da carteira de ações. Os potenciais benefícios para este setor incluem maior participação de clientes, melhor fidelização de clientes, aumento do faturamento e diminuição de riscos financeiros.

No setor de **saúde,** o uso do *Big Data* permite monitorar e prever o desenvolvimento de epidemias e surtos de doenças. Integrar dados de registros médicos com análise de mídias sociais permite-se monitorar surtos de gripe em tempo real, simplesmente por identificar o que as pessoas estão dizendo nas redes sociais.

No setor de **tecnologia**, a utilização do *Big Data* ajuda máquinas e dispositivos a se tornar mais inteligentes e autônomos. Por exemplo, ferramentas de *Big Data* são usadas para operar o automóvel auto-dirigido da Google. O Toyota Prius está equipado com câmeras, GPS, bem como computadores poderosos e sensores para conduzir com segurança na estrada sem a intervenção de seres humanos.

O setor de **óleo e gás** pode se beneficar de projetos de *Big Data* para minimizar o risco de acidentes de escavação e otimizar o processo de escavação a partir de análises de dados de sensores terrestres e dados geológicos.

No caso de **políticas públicas**, o *Big Data* contribui para o desenvolvimento das chamadas “cidades inteligentes”, por exemplo, permite às cidades otimizar o fluxo de tráfego com base em informações de tráfego em tempo real, bem como em mídia social e dados meteorológicos.

2.1 Características do *Big Data*

A necessidade de solucionar problemas reunindo e analisando dados de diversas naturezas, deu origem a pesquisas que nos levaram ao *Big Data*. Estas pesquisas foram desenhadas a partir de três aspectos iniciais (DAVENPORT, 2014):

- A múltipla natureza dos dados – aspecto relacionado com as diferentes fontes disponíveis;
- O uso de processamento em nuvem – aspecto relacionado ao uso ilimitado de recursos computacionais e com processamento em larga escala, com a possibilidade de redução de custos (economia de escala – é o aspecto econômico-financeiro);
- Uso de tecnologias específicas, tais como processamento de rotinas em paralelo e ferramentas para otimização como *Machine Learning* e *Analytics*.

Devido à possibilidade de capturar e analisar dados não estruturados, ou seja, imagens, vídeos, *logs*, *likes, tweets*, e a capacidade de fazer isto em tempo real, foram atribuídas características ao conceito consideradas os 5 V’s do *Big Data*: volume, variedade, velocidade, veracidade e valor.

Volume, como o próprio nome sugere, trata-se da quantidade de dados gerados pela internet como um todo. Segundo Cezar Taurion (2016), são gerados centenas de petabytes de dados por dia, e estima-se que este valor dobre a cada 18 meses. O impulso dado pela tecnologia, principalmente pelo aumento do uso dos dispositivos móveis, trouxe um forte incremento no volume de dados (RIBEIRO, 2014, p. 97).

Heath e Bizer (2011) reforçam que na atualidade estamos cercados por uma grande quantidade de dados e informação. São registros sobre o cotidiano, desempenho da educação, produção de bens e serviços, investimentos, impostos governamentais, estatísticas sobre a economia e dados sobre o consumo que nos ajudam a tomar decisões e gerar conhecimento.

Tabela 1 – Grandezas

Unidade	Símbolo	Valor Equivalente	Múltiplo
Bit	b		
Byte	B	8 bits	B = 10^0 byte
Kilobyte	KB	1024 B	KB =10^3 byte
Megabyte	MB	1024 KB	MB = 10^6 byte
Gigabyte	GB	1024 MB	GB = 10^9 byte
Terabyte	TB	1024 GB	TB = 10^12 byte
Petabyte	PB	1024 TB	PB = 10^15 byte
Exabyte	EB	1024 PB	EB = 10^18 byte
Zettabyte	ZB	1024 EB	ZB = 10^21 byte
Yottabyte	YB	1024 ZB	YB = 10^24 byte

Fonte: Tecmundo (2009).

Segundo a New England College (2017), em 1992 eram gerados 100 GB por dia de dados no tráfico global da rede. Em 1997, aumentou-se para 100 GB por hora; 2002 100 GB por segundo; em 2007 o tráfico chegou à casa dos 2 terabytes (TB) por segundo; em 2014, 16 TB por segundo e a previsão é de que em 2019 este valor chegue à casa dos 50 TB por segundo de dados gerados por todo tráfico na rede. A tabela 1 acima ajuda a representa a grandeza dos valores. No ano de 2016 registrou-se 1.1 zettabytes (ZB) de dados registrados.

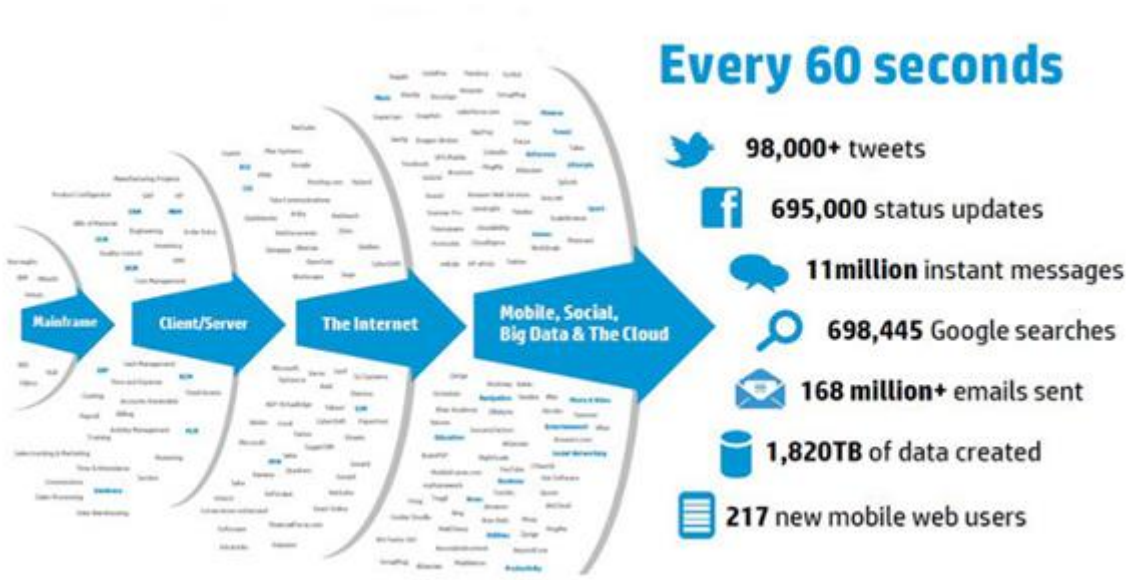
Variedade diz respeito às diferentes fontes de informações de onde os dados são gerados, na qual a maioria provém de fontes de dados não estruturados, como e-mails, mídias sociais (facebook, linkedin, twitter, youtube), sensores, câmeras de vídeos, entre outros. Apenas 1% destes dados é efetivamente analisado (BREITMAN, 2013). O uso da informação pela sociedade tem se modificado ao longo do tempo e, consequentemente, surgem novos modelos econômicos e tecnológicos. A utilização de dispositivos móveis como meio de comunicação teve um aumento significativo, e o uso cada vez maior da Internet vem ultrapassando as

barreiras que encontrávamos para nos comunicar, e ao mesmo tempo demarcando novos limites para a sociedade contemporânea (RIBEIRO, 2008, p. 15).

Qualquer cidadão pode gravar e postar um vídeo em mídias sociais ou no *Youtube*. Estima-se que a quantidade de vídeos produzidos diariamente ultrapassa a produção dos primeiros 50 anos de televisão (DAVENPORT, 2014).

A figura abaixo ilustra as mais diversas fontes gerando grandes quantidades de dados, o que representa os dois primeiros V's.

Figura 2: Mundo dos dados



Fonte: http://targetrichsolutions.com

Velocidade está relacionada com a agilidade que se precisa ter sobre os dados que estão sendo capturados em tempo real, como, por exemplo, o controle de tráfego nas ruas. As melhorias de canais de transmissão, como redes de fibra ótica, bandas de telefonia celular e emissores de canais de alta capacidade contribuem para atingir uma maior velocidade na troca de dados e informações (MATTOSO, 2013).

Segundo Florissi (2012), o desenvolvimento de tecnologia de processadores e discos para armazenamento duplica de poder a cada dois anos, sinal de que a velocidade continuará aumentando.

Veracidade trata-se da qualidade dos dados e informações geradas a partir destes para que façam sentido e tragam informações autênticas que contribuam para uma correta tomada de decisão ao final de um estudo.

Por fim, valor, quanto maior a riqueza de dados, mais importante é saber realizar as perguntas certas no início de todo processo de análise (Brown, 2014). É necessário estar focado para a orientação do negócio, o valor que a coleta e a análise dos dados trará para o mesmo. Não é viável realizar todo o processo de *Big Data* se não se tem questionamentos que ajudem o negócio de modo realístico. Da mesma forma é importante estar atento aos custos envolvidos nessa operação, o valor agregado de todo esse trabalho desenvolvido, coleta, armazenamento e análise de todos esses dados tem que compensar os custos financeiros envolvidos (Taurion, 2013).

2.2 As Fases do *Big Data*

O processo de um projeto de *Big Data* é realizado por meio de cinco atividades que precisam ser conduzidas com um propósito bem definido, ou seja, o problema ou oportunidade que está sendo abordado necessita ter metas e objetivos claros. A primeira tarefa trata da obtenção dos dados, identificando quais são releventes para o estudo realizado e transportando-os para bases de dados que serão utilizadas. Primeiramente, identifica-se todas as fontes de dados, podendo ser de diversos tipos de arquivos, bancos de dados, internet, dispositivos móveis, satélites, entre outras. Após a identificação dos dados e suas fontes, o próximo passo é coletar os dados e integrar as diversas fontes. Nesta etapa pode ocorrer a conversão do formato dos dados, uma vez que provém de fontes distintias.

Uma vez que os dados estejam coletados e integrados e tenham coerência para análise, o próximo passo é a preparação. Esta etapa é dividida em dois passos: exploração e pré-processamento. Na exploração é realizada uma investigação preliminar com o intuito de obter melhor entendimento das características dos dados, orientando o resto do processo. Com a exploração, será buscado, inicialmente, técnicas de correlação, tendências gerais, *outliers*, entre outras. A correlação irá

fornecer informações a respeito do relacionamento entre as variáveis nos dados. Tendências gerais irão revelar se as variáveis estão se movendo para alguma direção específica, como o aumento do volume de transações bancárias durante o ano. *Outliers* ajudam a identificar possíveis problemas com os dados, ou mesmo indicar um ponto que necessita de envolvimento analítico mais profundo. De forma geral, sem a fase de exploração não é possível utilizar os dados em sua forma mais eficaz. Uma maneira de explorar seus dados é calcular estatísticas resumidas para descrever numericamente os dados. Estatísticas resumidas capturam várias características de um conjunto de valores em um único número. Algumas estatísticas de resumo básica que devem ser calculadas no conjunto de dados são: média, mediana, moda e desvio padrão. Média e mediana são medidas da localização de um conjunto de valores. Moda é o valor que ocorre mais freqüentemente em seu conjunto de dados, e desvio padrão é uma medida de dispersão em seus dados. Olhando para essas medidas será possível ter ideia da natureza dos dados, podendo dizer se há algo errado com os mesmos. Por exemplo, se o intervalo dos valores de idade inclui números negativos ou um número muito maior que cem, há algo suspeito nos dados que precisam ser examinados. (May Hyugan, 2016).

Após a exploração dos dados, inicia-se a pré-preparação para análises futuras. As principais atividades nesta etapa é a limpeza dos dados, seleção de variáveis apropriadas e transformação dos dados conforme necessário. Uma parte muito importante da preparação de dados é a limpeza para resolver problemas de qualidade. Problemas de qualidade com dados incluem valores em falta ou dados duplicados, como dois registros diferentes para o mesmo cliente com endereços diferentes, ou CEP inconsistente com nove dígitos.

Durante a etapa de exploração de dados, pode ter sido descoberto, por exemplo, que dois recursos são muito correlacionados. Nesse caso, um desses recursos pode ser removido sem afetar negativamente os resultados da análise. A remoção de recursos redundantes ou irrelevantes tornará a análise subsequente mais simples.

Após os dados terem sido preparados, a fase de análise é iniciada com a seleção de técnicas de *machine learning* (aprendizagem de máquina), e então

construir o modelo com os dados previamente preparados. Estas técnicas serão abordadas posteriormente neste trabalho.

O ponto central desta análise está ligado à capacidade de correlacionar dados, pois, em essência, quando temos uma pequena quantidade de dados não temos muita dificuldade de correlacioná-los, pois existem poucas inter-relações. Mas, com uma grande quantidade, temos muitos dados sendo gerados em paralelo, logo, surgem dificuldades de correlacioná-los (SEYMOUR, 2014).

Segundo Sathi (2013), é em busca de uma melhor capacidade de correlacionar dados que processos de negócios das empresas, governos e outras instituições começam a se integrar, visando a uma mudança de comportamento dos executivos e na ótica de produção de bens e serviços que influencia estas organizações. Todo este trabalho e processo ganha o nome de *Big Data Analytics*, o qual criou um novo perfil profissional denominado *Data Scientist* (Cientista de Dados). A característica principal deste profissional é ter a capacidade de aplicar ferramentas analíticas e algoritmos para gerar previsões sobre produtos, serviços, e comportamento de indivíduos (DAVENPORT; PATIL, 2012, p. 70-76). Este profissional deve ter conhecimento em matemática e estatística, aliado a estratégias para tratamento de grandes conjuntos de dados, fazendo uso de modelos matemáticos, formulação de hipóteses e técnicas de regressão com os experimentos realizados e respostas obtidas.

Segundo Marchand e Peppard (2013), o objetivo de projetos de *Big Data* é a busca do desenvolvimento de novos produtos, compreender necessidades e comportamentos dos clientes, bem como perceber novos mercados. Também afirmam que é necessário construir hipóteses e identificar dados e informações relevantes para tratar com clientes e usuários. Completam propondo que este processo deve ser repetido e refinado de acordo com as experiências e resultados obtidos. Há também objetivos direcionados às políticas públicas, como mitigar a fraude, na aplicação da lei para prever futuros crimes, na educação para melhorar o desempenho acadêmico e na infraestrutura para monitorar o uso de recursos preciosos nas cidades. (Datafloq, 2017).

O tratamento de dados é uma maneira de tornar visível o que antes era invisível, ou estava oculto. Para isso, são utilizados ferramentas e métodos

computacionais avançados, como o processo de mineração de dados, que será aprofundado no capítulo seguinte.

3 MINERAÇÃO DE DADOS

Uma das etapas do *Big Data* é constituída da mineração dados, ou *Data Mining.* A mineração de dados contempla ferramentas e técnicas de análise que verifica dentro de grandes volumes de dados se há alguma tendência ou agrupamento, muitas vezes implícito. Segundo Grilo Júnior (2010), o processo de mineração de dados utiliza fórmulas matemáticas e estatísticas e técnicas avançadas, como redes neurais, que têm como característica a habilidade de aprender com o seu próprio ambiente e assim melhorar o desempenho, técnicas heurísticas para se resolver um determinado problema quando não se sabe se a solução está correta, e descobertas por regra de detecção de desvio.

Giudici (2003) define *Data Mining* como a seleção, exploração e modelagem de grande volume de dados para descobrir relações e padrões desconhecidos ou empíricos, objetivando resultados consistentes e úteis a partir de um banco de dados.

Ainda segundo Giudici (2003), diferente de relatórios e consultas, onde os relacionamentos já se conhecem, a função da mineração de dados é desvendar o que não se sabe sobre os dados armazenados em um banco de dados. Um exemplo clássico e prático de aplicação de *Data Mining*, é a utilização dos dados de vendas em estabelecimentos varejistas para descobrir supostas relações entre produtos sem conexão aparente, mas que muitas vezes são vendidos juntos. Outro exemplo, citado pela Datafloq (2017), é o uso de algoritmos para comparar registros públicos com o imposto de renda, a fim de detectar inconsistências. Segundo o site, em 2014 o estado de Indiana, nos EUA, identificou aproximadamente 75.000 retornos fraudulentos que resultado em uma economia de $ 85 milhões de dólares. 62% de governos locais dos Estados Unidos já usam ou estão em processo de iniciar a utilização de *Big Data* para identificar fraudes do imposto de renda, uma vez que é

uma tarefa difícil de ser realizada e tem contribuído para o roubo de bilhões de dólares registrados anualmente.

Data Mining tem o propósito de extrair conhecimento onde para um observador humano seria quase impossível, devido a sua dimensão, complexidade e volume de dados (Dutra, 2005).

A mineração de dados busca identificar padrões através de técnicas dentro de um grande volume de dados, revelando detalhes implícitos ou mesmo empíricos, não comprovados.

A mineração de dados é o resultado da compatibilização da estatística convencional com técnicas de inteligência artificial a fim de converter os dados em informação. Contudo, Barbieri (2001) ressalta que “as informações geradas pelas ferramentas de *Data Mining* estão ligadas com o tratamento da informação, e não com a estruturação dos dados”.

O’Brian (2004) reforça que os *softwares* de *Data Mining* utilizam algoritmos bastante elaborados de reconhecimento de padrões, aliados a uma diversidade de técnicas matemáticas e estatísticas para observar um grande volume de dados e extrair informações relevantes, úteis e estratégicas até então desconhecidas.

3.1 Ferramentas de Mineração de Dados

Esta seção apresenta ferramentas utilizadas no mercado para mineração de dados. A tabela 2 apresenta alguns *softwares* utilizados para a mineração de dados, bem como as principais tarefas e seus respectivos fabricantes.

Tabela 2 - Ferramentas de Mineração de Dados

Ferramenta	Tarefas	Fabricante
Oracle Data Miner	Classificação Regressão Associação Clusterização Detecção de Anomalia	Oracle www.oracle.com
PolyAnalyst	Regressão Associação Clusterização Sumarização	Megacomputer Intelligence www.megacomputer.com
SPSS/Clementine	Classificação Associação Clusterização	SPSS Inc. www.spss.com
KNIME	Classificação Regressão Associação Clusterização	KNIME www.knime.org
WizSoft	Sumarização Clusterização Detecção de Anomalia	WizSoft Inc. www.wizsoft.com
SAS Enterprise Miner	Classificação Associação Clusterização Agrupamento	SAS Corp. www.sas.com

Fonte: Elaboração própria

3.2 Algoritmos de Mineração de Dados

Nesta parte será abordado brevemente as principais tarefas utilizadas na mineração de dados (classificação, regressão, associação, clusterização, detecção de anomalia) e seus respectivos algoritmos, que são disponibilizados no *software* da empresa Oracle conforme a tabela 3 na página 25.

A tabela 3 ilustra as tarefas do *Oracle Data Miner* e seus respectivos algoritmos utilizados para mineração. Cabe ressaltar que o mesmo algoritmo, utilizado em diferentes tarefas, pode ter seu parâmetro variado significativamente, dependendo do projeto. Além disso, existem outros algoritmos que podem ser utilizados nas tarefas listadas na tabela 3, mas para este trabalho, usaremos os

disponibilizados no *software* da empresa Oracle, o *Oracle Data Miner*, conforme mencionado anteriormente.

Tabela 3 – Algoritmos de Mineração de Dados

Tarefa	Algoritmos de Mineração
Classificação	SVM (Support Vector Machine) GLM (Generalized Linear Model) Árvore de Decisões Naive Bayes
Regressão	SVM (Support Vector Machine) GLM (Generalized Linear Model)
Associação	Apriori
Clusterização	K-Means O-Cluster
Detecção de Anomalia	SVM (Support Vector Machine)

Fonte: Data Science Academy (2015)

No capítulo 4 deste trabalho serão abordados os conceitos das tarefas listadas na tabela acima e em seguida, no capítulo 5 deste trabalho, utilizaremos os algoritmos da tarefa de classificação com o objetivo de prever a aquisição de cartões fidelidade de um supermercado por parte dos clientes baseado nas suas características pessoais obtidas durante uma campanha realizada.

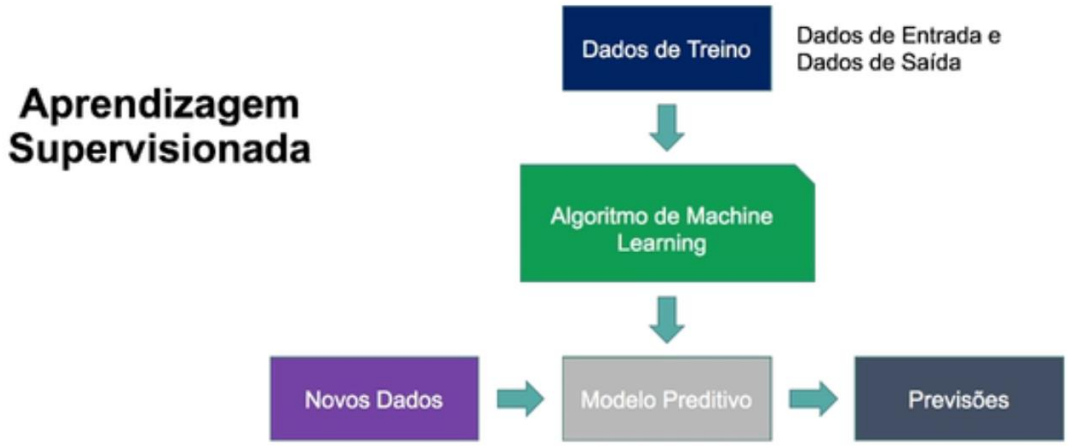
Segundo Fayyad (1996), a escolha do algoritmo faz parte do processo inicial de mineração, onde se define qual tarefa será executada e qual o tipo de informação o algoritmo deverá extrair, portanto, não existe uma forma de mineração de dados genérica.

3.3 Aprendizagem Supervisionada e Não Supervisionada

Quando tratamos de aprendizagem de máquina, estamos tratando de algoritmos que aprendem a partir dos dados e geram modelos preditivos, representados por funções matemáticas. O processo de aprendizagem dos algoritmos compreende dois grupos principais:

- Aprendizagem supervisionada;
- Aprendizagem não supervisionada;

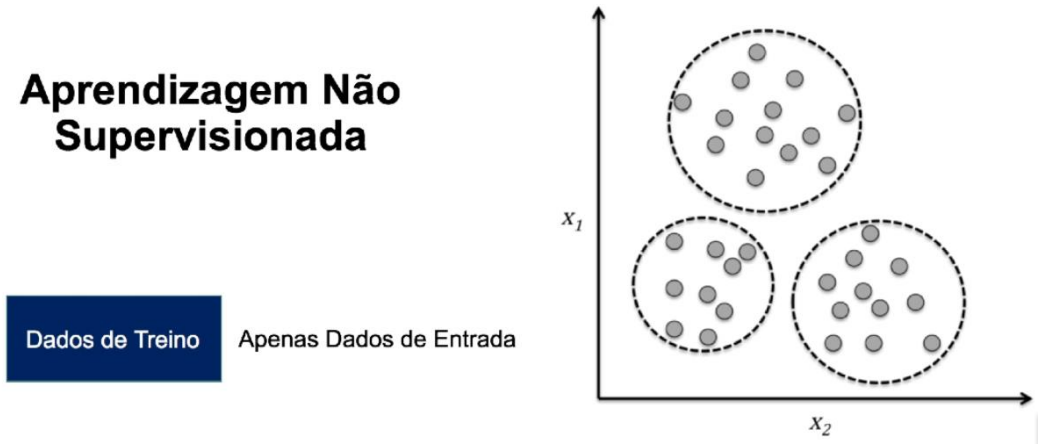
Figura 3: Aprendizagem Supervisionada



Fonte: Data Science Academy

As tarefas que tem como característica a aprendizagem supervisionada possuem variáveis de saída, também chamadas de variáveis *target* ou variáveis dependentes, que são previstas por um grupo de variáveis preditoras, ou variáveis independentes. Utilizando este grupo de variáveis, gera-se uma função que mapeia entradas para saídas desejadas. O processo de treinamento continua até o modelo atingir um nível de precisão desejável nos dados de teste. Os conceitos de dados de treino e dados de teste serão abordados com maior profundidade nas próximas páginas deste trabalho. Exemplos de aprendizagem supervisionada: regressão, árvore de decisão, *random forest.*

Figura 4: Aprendizagem não supervisionada



Fonte: Data Science Academy.

Na aprendizagem não supervisionada, por sua vez, não há variável de saída a ser estimada. Neste caso, são feitos agrupamentos na população de dados em diversos grupos *(clusters)*, utilizados para segmentar as observações com caracteristicas específicas. A partir dos dados de entrada apresentados ao algoritmo, cabe a este descobrir as interrelações dos dados. Exemplos de aprendizagem não supervisionada são: clusterização e análise de associação (apriori).

4 MODELOS DE MINERAÇÃO DE DADOS

Este capítulo irá abordar os conceitos de classificação, regressão, associação, clusterização e detecção de anomalia, listados na tabela 3 da página 22 e exemplificar como estes modelos são usados para resolver problemas no mundo real.

4.1 Classificação

A **classificação** é um método supervisionado e é uma das técnicas mais comuns de mineração de dados e amplamente utilizada em problemas em que os dados históricos são rotulados para indicar se um evento específico aconteceu ou não. Este conjunto de dados pré-rotulados é chamado de **conjunto de dados de treinamento**. Este conjunto de dados de treinamento consiste nos dados para os quais o resultado já é conhecido. Por exemplo, ao realizar a análise de crédito de um determinado cliente, utilizaria-se um conjunto de dados de todos os clientes que já passaram por essa análise de crédito anteriormente. Este conjunto de dados de treinamento contém um atributo resposta para cada cliente, e é este atributo ao qual denominamos variável alvo, ou variável *target.* Esta variável, por sua vez, que contém o rótulo, normalmente binário (0 ou 1) e categórico, que é usado pelos algoritmos de classificação para construir os modelos. Apresentam-se ao algoritmo os dados de entrada, ou seja, os atributos e características da entidade que está sendo pesquisada, e a variável de resposta, que é o objetivo que se deseja alcançar, tudo baseado em dados históricos. Em seguida, constrói-se o modelo que será capaz de receber novos dados de entrada e prever a variável alvo para novos clientes. Exemplos de aplicação de modelos de classificação:

- Determinar o diagnóstico de uma doença em um paciente, observando as características similares em outros grupos de pacientes;

- Previsão se uma pessoa irá gostar da recomendação de filmes ou músicas;

- Identificação de padrões em dados genéticos, detectando doenças específicas.

Segundo Mendes (2012), o processo de classificação é dividido em três etapas. Primeiro, construir o modelo, identificando o problema a ser resolvido, buscando os dados que irão dar suporte à análise, realizar algum procedimento de pré-processamento caso necessário e, então, apresentar os dados aos algoritmos, e este algoritmo irá buscar uma função matemática que estabelece o melhor relacionamento entre os dados. Nesta etapa, seleciona-se um subconjunto de dados para a construção do modelo (estes são os dados do conjunto de dados de

treinamento). O resultado desta operação da apresentação dos dados ao algoritmo irá gerar o modelo preditivo.

Em segundo lugar, é necessário testar o modelo. Uma vez que o modelo esteja criado, testa-se o modelo apresentando um novo conjunto de dados, chamados de **dados de teste.** Isto permite medir a acurácia, ou nível de precisão, do modelo. Nota-se que neste passo, pode ser utilizado o mesmo conjunto de dados que foi utilizado para os dados de treinamento, porém divididos anteriormente em, por exemplo, 70% para dados de treinamento e 30% para dados de teste. O motivo de não se utilizar 100% dos dados para treinamento é para que não ocorra *overfitting*, que é basicamente o algoritmo decorar os resultados e não medir a acurácia de forma precisa para os novos dados, esperando sempre as mesmas características utilizadas para aprender.

Por fim, aplica-se o modelo, colocando-o em produção para que o problema identificado na primeira fase seja resolvido com a introdução de um novo conjunto de dados.

O tipo mais simples de classificação é a classificação binária, na qual o atributo de destino possui apenas dois valores possíveis, por exemplo, uma classificação de risco de crédito é considerada alta ou baixa.

Na classificação multi-classe, é possível ter como saída mais de dois valores. Conforme exemplo anterior, uma situação de risco de crédito poderia ser baixa, média ou alta.

O modelo de classificação será usado posteriormente neste trabalho para resolver um problema de previsibilidade de cartões fidelidade.

Os principais algoritmos utilizados na técnica de classificação, conforme a tabela 3, serão melhor aprofundados no próximo capítulo durante o estudo de caso.

4.2 Regressão

A **regressão**, assim como a classificação, é uma técnica supervisionada utilizada para prever uma saída numérica com base em outros atributos e seu conjunto de dados. Essa saída numérica pode ser uma variável de valor contínuo, diferente do atributo de destino usado na classificação, na qual é prevista uma saída com uma variável categórica. Exemplos de onde a regressão pode ser usada são:

- Previsão de compras de clientes;

- Cálculo de dosagens para prescrições médicas;

- Previsões financeiras com base no histórico;

- Limites de crédito.

A regressão em *data mining* é uma técnica de mineração de dados usada para prever um valor contínuo.

Um exemplo de aplicação desta técnica é coletar dados históricos de compras de diversos clientes e a partir da análise de regressão, é possível prever quanto um cliente poderá gastar no mês seguinte, baseados em uma ou mais variáveis de entrada.

A regressão dá suporte ao entendimento de como determinadas variáveis influenciam outra variável. O objetivo da regressão é prever uma saída que seja representada por um valor numérico. Outro exemplo da utilização da regressão é calcular o valor do tempo de vida de um cliente de uma empresa. Para construir um modelo de regressão, são necessários dados históricos de compras dos clientes e diversos outros atributos, como: idade, renda, estado civil, profissão, entre outros. Estes dados são apresentados ao algoritmo de regressão e, então, será gerado um modelo que será capaz de prever quanto um cliente irá gastar em produtos ou serviços enquanto ele for cliente da empresa.

O processo de construção, aplicação e teste de um modelo de regressão, segue o mesmo processo utilizado na classificação, conforme visto no item 3.2.1 deste trabalho. Utiliza-se o conjunto de dados que possui o atributo de valor

contínuo que será previsto e este valor já está determinado através de dados históricos. Os algoritmos de regressão dividem este conjunto de dados em conjuntos de dados de treino e de teste. O conjunto de dados de treinamento é usado para treinar e construir o modelo de regressão. Em seguida, aplica-se o modelo de regressão ao conjunto de dados de teste. O modelo prediz o valor da variável contínua e, em seguida, compara o valor previsto com o valor real do conjunto de dados. Com isso, é possível calcular a precisão dos modelos de regressão.

Para exemplificar ambos os modelos supervisionados, classificação e regressão, em um problema de aprovação de crédito a um indivíduo, utiliza-se a classificação para a decisão de oferecer o crédito (sim/não). Em seguida, utiliza-se a regressão para definir a quantidade em dinheiro que será disponibilizado ao indivíduo.

Modelos de regressão são modelos matemáticos que relacionam o comportamento de uma variável *y* com outras variáveis *x*. A variável *x* é a variável independente da equação, enquanto *y* é dependente das variações de *x.* Um modelo de regressão pode ser simples, quando envolve a relação entre duas variáveis, ou múltiplo, quando envolve uma relação com mais de duas variáveis.

$$\hat{Y}_i = \alpha + \beta . X_i \quad \text{(Simples)}$$

$$Y_i = a + b_1 . X_{1i} + b_2 . X_{2i} + \ldots + b_k . X_{ki} + u_i \quad \text{(Múltiplo)}$$

A variável *x,* que são os atributos, ajuda a explicar a variável *y,* a variável de saída que se deseja prever.

A regressão será um modelo utilizado posteriormente no capítulo 5 deste trabalho como forma de dar suporte à análise preditiva do PIB brasileiro.

4.3 Análise de Associação

A **análise de associação** é uma técnica de mineração de dados não supervisionada que permite a associação entre itens que fazem parte de um evento. Esta técnica é comumente aplicada em análise de cestas de mercado (*market basket analysis), cross-selling,* ou venda cruzada em produtos financeiros, análise de sinistro de seguro, entre outras. (MENDES, 2015).

Segundo os autores Goldschmidt e Passos (2005), a tarefa de análise de associação busca encontrar elementos que aconteçam de forma frequente e simultânea no banco de dados.

Gonçalves (2005) exemplifica a utilização deste tipo de algoritmo na análise das transações de compras, onde se verifica padrões de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto.

Contudo, as associações de produtos em cestas de compras não devem considerar apenas associações triviais, como, por exemplo, quem compra leite também compra pão, mas sim aquelas que não são óbvias e que podem se tornar importante fonte de informação na tomada de decisão. (SILVEIRA, 2003).

Um exemplo de resultado de uma análise de associação em uma loja de departamento pode ser representado como: “68% dos clientes que compram o produto ‘calça jeans’ também compram o produto ‘camisa de malha branca’. 5,5% de todas as compras contém os dois produtos.”

Neste exemplo, calça jeans seria o produto considerado antecedente e a camisa de malha o produto consequente.

4.4 Clusterização

A **clusterização,** também denominada análise de agrupamento, é o processo de dividir os dados em pequenos grupos, denominados *cluster.* Dentro de cada

cluster os dados são semelhantes entre si, mas apresentam diferenças com dados de outros *clusters.* “É um método de segmentação de dados que partilham tendências e padrões semelhantes.” (MENDES, 2015). É também considerada uma técnica de aprendizagem não supervisionada, pois não se busca o perfil de uma peculiariedade específica.

Nesta técnica, não é necessário definir os grupos nem os atributos que deverão ser utilizados para segmentar o conjunto de dados. É, também, considerada um dos primeiros passos a realizar em um estudo de mineração de dados, uma vez que identifica grupos que podem ser utilizados como ponto de partida para mais explorações de relações.

O problema de clusterização é do interesse de qualquer área que se deseja agrupar dados, por exemplo:

- Compras efetuadas em um supermercado;

- Especificações físicas e químicas de petróleo;

- Sintomas de doenças;

- Características de seres vivos;

- Transações bancárias realizadas por clientes de um banco.

Em resumo, clusterização é útil para redução de dados, reduzindo uma grande quantidade de dados para um número de subgrupos característicos, permitindo o desenvolvimento de esquemas de classificação e sugerindo ou apoiando hipóteses sobre a estrutura dos dados.

Um dos principais algoritmos de clusterização é o *k-means.* Este algoritmo utiliza medida de distância baseada em densidade, e pode tratar de grupos de dados de qualquer tamanho, mas com baixo número de atributos.

Os modelos são construídos de forma hierárquica, utilizando uma abordagem *top-down*, com divisão binária e refinamento de todos os nós, que são os agrupamentos de dados, até o último disponível. O centróide dos nós mais internos na hierarquia é atualizado para refletir as mudanças à medida que o algoritmo vai avançando e categorizando os dados. O nó com maior variância é dividido para

aumentar o tamanho de *clusters*, até que o número de *clusters* previamente definido (a letra *k* representa o número de clusters), seja atingido. Portanto, o nó com maior variância é dividido, gerando dois sub-nós, então o algoritmo analisa cada um destes sub-nós, verificando mais uma vez a variância, e se encontrar uma alta variância, subdivide e assim sucessivamente.

4.5 Detecção de Anomalia

A **detecção de anomalia** é a busca e identificação de registros de casos que não estejam de acordo com um padrão típico. Estes casos não conformes são referidos como anomalias ou *outliers.* (MENDES, 2015).

O objetivo da detecção de anomalia é identificar casos que são incomuns dentro de um conjunto de dados aparentemente homogêneos. A detecção de anomalia é uma ferramenta importante para a detecção de fraude, intrusão de rede e outros eventos raros que podem ter um grande significado, mas são difíceis de detectar, e podem ser usadas para resolver problemas como:

- Identificação de reivindicações fraudulentes de seguro;

- Monitoramento de transações financeiras que requerem investigação adicional para operações potencialmente fraudulentas;

- Monitoramente de rede para identificar a presença de *hackers.*

5 APLICAÇÕES DE MINERAÇÃO DE DADOS

Com cada vez mais dispositivos eletrônicos conectados, como smartphones, laptops, sensores e veículos, novas conexões surgem em todo lugar. Estas conexões geram um amontoado de dados, e estes dados representam uma oportunidade como fonte de vantagem competitiva para as empresas que conseguem gerenciá-los e transformá-los para análise.

A utilização do *Big Data* pode ser de grande valia em boa parte das indústrias, empresas e governos que usam a mineração de dados como uma etapa para extrair a novas informações do cliente e usar isto a favor para gerar novas vendas, otimizar a produção, melhorar a logística, diminuir custos e realizar melhorias na questão das políticas públicas de forma geral.

Alguns exemplos que ilustram o uso do *Big Data* são:

- Compradores podem receber descontos diretamente no smartphone ao se aproximarem ou entrarem em alguma loja;
- Consumidores online recebem ofertas personalizadas exclusivas, por exemplo em redes de supermercados como Pão de Açucar, ou da indústria farmacêutica, como Drogarias Raia;
- Fabricantes medem o sucesso de seus novos produtos diariamente, ao invés de semana;
- Motoristas podem utilizar recursos em tempo real que dão acesso à informação de rotas com menor tráfego através de aplicativos de GPS, como o Waze;
- Reconhecimento de Imagens: é aplicado para identificar caracteres escritos, comparar e identificar rostos, aplicação de filtros de equipamentos fotográficos e detectar comportamentos suspeitos, por meio de câmeras de segurança;
- Implementar o policiamento preditivo, o qual envolve algoritmos e dados de tipo, local e tempo de crimes cometidos anteriormente, a fim de atribuir probabilidades de eventos de criminalidade futuros a regiões de espaço e horário.
- A mineração de dados e a análise preditiva podem ser usadas para identificar os estudantes com risco de abandonar os estudos. O monitoramento das taxas de retenção dos alunos possibilitará melhorar o desempenho acadêmico dos alunos e, portanto, satisfação geral entre estudantes, professores e a administração.

5.1 Aplicações Técnicas de Mineração de Dados

Para exemplificar o uso de uma das técnicas para modelagem de previsão citadas anteriormente neste trabalho, utiliza-se uma ferramenta da empresa Oracle denominada Oracle *Data Miner*. O objetivo é demonstrar, a partir de uma tabela de dados fictícia, a utilização e aplicabilidade em benefício aos usuários.

Para a modelagem de previsão, considere uma amostra da tabela 4 utilizada, consultada diretamente da base de dados onde estão os dados disponíveis*:

Tabela 4 - Amostra de dados hipotéticos de clientes de uma empresa fictícia

	CUST_ID	AFFINITY_CARD	AGE	CUST_GENDER	EDUCATION	CUST_INCOME_LEVEL	COUNTRY_NAME
1	101,501.0000	0.0000	41.0000	F	Masters	J: 190,000 - 249,999	United States of America
2	101,502.0000	0.0000	27.0000	M	Bach.	I: 170,000 - 189,999	United States of America
3	101,503.0000	0.0000	20.0000	F	HS-grad	H: 150,000 - 169,999	United States of America
4	101,504.0000	1.0000	45.0000	M	Bach.	B: 30,000 - 49,999	United States of America
5	101,505.0000	1.0000	34.0000	M	Masters	K: 250,000 - 299,999	United States of America
6	101,506.0000	0.0000	38.0000	M	HS-grad	K: 250,000 - 299,999	United States of America
7	101,507.0000	0.0000	28.0000	M	< Bach.	J: 190,000 - 249,999	United States of America
8	101,508.0000	0.0000	19.0000	M	HS-grad	K: 250,000 - 299,999	United States of America
9	101,509.0000	0.0000	52.0000	M	Bach.	K: 250,000 - 299,999	Brazil
10	101,510.0000	1.0000	27.0000	M	Bach.	L: 300,000 and above	United States of America
11	101,511.0000	0.0000	30.0000	M	Bach.	H: 150,000 - 169,999	United States of America
12	101,512.0000	0.0000	30.0000	F	Profsc	I: 170,000 - 189,999	United States of America
13	101,513.0000	0.0000	31.0000	M	Bach.	J: 190,000 - 249,999	United States of America
14	101,514.0000	0.0000	45.0000	M	HS-grad	L: 300,000 and above	United States of America
15	101,515.0000	0.0000	36.0000	F	11th	J: 190,000 - 249,999	United States of America
16	101,516.0000	0.0000	33.0000	M	< Bach.	G: 130,000 - 149,999	United States of America
17	101,517.0000	0.0000	38.0000	F	HS-grad	I: 170,000 - 189,999	United States of America
18	101,518.0000	0.0000	22.0000	M	5th-6th	L: 300,000 and above	Argentina
19	101,519.0000	0.0000	46.0000	F	< Bach.	J: 190,000 - 249,999	Brazil
20	101,520.0000	1.0000	39.0000	M	HS-grad	B: 30,000 - 49,999	United States of America
21	101,521.0000	0.0000	61.0000	M	HS-grad	L: 300,000 and above	United States of America
22	101,522.0000	1.0000	39.0000	F	Masters	J: 190,000 - 249,999	United States of America
23	101,523.0000	0.0000	22.0000	M	HS-grad	L: 300,000 and above	United States of America
24	101,524.0000	0.0000	38.0000	M	HS-grad	I: 170,000 - 189,999	United States of America
25	101,525.0000	0.0000	18.0000	F	HS-grad	K: 250,000 - 299,999	United States of America
26	101,526.0000	1.0000	40.0000	M	Profsc	I: 170,000 - 189,999	United States of America
27	101,527.0000	0.0000	19.0000	M	< Bach.	J: 190,000 - 249,999	United States of America
28	101,528.0000	0.0000	52.0000	M	HS-grad	K: 250,000 - 299,999	United States of America
29	101,529.0000	0.0000	22.0000	M	< Bach.	K: 250,000 - 299,999	United States of America
30	101,530.0000	1.0000	34.0000	M	Bach.	H: 150,000 - 169,999	United States of America
31	101,531.0000	0.0000	28.0000	M	< Bach.	J: 190,000 - 249,999	Argentina

Fonte: Data Science Academy

A tabela 4 acima que será utilizada como exemplo de um modelo de classificação apresenta um conjunto de dados históricos de clientes de um mercado de varejo, com as seguintes informações nas colunas:

CUST_ID: é o número identificador único de cada registro dos clientes, onde contém as informações pessoais na linha;

AFFINITY_CARD: esta coluna indica se o cliente possui ou não cartão fidelidade, onde o número 0 indica que não possui e 1 indica que possui;

AGE: coluna que indica a idade do cliente;

EDUCATION: esta coluna indica o grau de escolaridade do cliente;

CUST_INCOME_LEVEL: esta coluna informa a faixa de renda em que o cliente se encontra;

COUNTRY NAME: coluna que indica o país de origem do cliente;

O objetivo desta tarefa é, a partir da tabela de dados históricos (Tabela 4), criar e treinar um algoritmo do modelo de classificação que poderá, futuramente, ser utilizado em uma nova base de dados e prever quais clientes possuem chances de adquirir o cartão fidelidade da empresa, baseado nas características de clientes que já o possuem.

Utiliza-se o modelo de classificação, no qual dados históricos são rotulados para indicar que um evento específico aconteceu ou não, por exemplo, indicar se um cliente possui cartão fidelidade ou não. É uma técnica supervisionada, ou seja, possui uma variável-alvo que contém um rótulo, normalmente binário e qualitativo, que é usado pelos algoritmos para construir os modelos. Em seguida, são apresentados os dados de entrada aos algoritmos, ou seja, os atributos da entidade que está sendo pesquisado, e a variável de resposta, que é o objetivo que se deseja alcançar, baseado na análise dos dados históricos. O modelo construído será capaz de receber novos dados de entrada e prever a variável alvo para novas situações. A classificação pode ser usada em diferentes áreas para diversas situações, como: reconhecimento de imagem e facial, previsão de recomendação de filmes ou músicas, identificação de padrões em dados genéticos para detectar doenças específicas, análise de crédito, prever decisão de compra, entre outras. (MENDES, 2013). O processo de classificação é dividido em três etapas:

1) Construir o modelo, a qual envolve identificar o problema, buscar os dados que irão suportar a análise, realizar um procedimento de limpeza dos

dados e, por fim, apresentar os dados a algum algoritmo, que por sua vez irá buscar uma função matemática que estabelece o melhor relacionamento entre os dados.

2) Testar o modelo, apresentando dados diferentes daqueles que foram usados durante o processo de treinamento para criar o modelo. Desta forma, permite-se medir a acurácia do modelo.

3) Por fim, aplica-se o modelo para resolver o problema ao qual ele foi designado.

Para ilustrar como funciona a construção deste modelo de classificação, usaremos uma base de dados com dados históricos de clientes de um mercado de varejo para prever o efeito de uma campanha de marketing com o objetivo de atrair novos clientes, mostrar ao mercado os produtos oferecidos e conseguir aumentar o volume de vendas, consequentemente, aumentando o faturamento. A empresa que realizou a campanha tinha como principal objetivo fornecer cartões fidelidade aos clientes, a fim de elaborar campanhas customizadas e níveis percentuais específicos de desconto. O objetivo desta parte é prever, com base na campanha realizada, o efeito da próxima campanha, quantos clientes irão adquirir cartão fidelidade e quantos não vão.

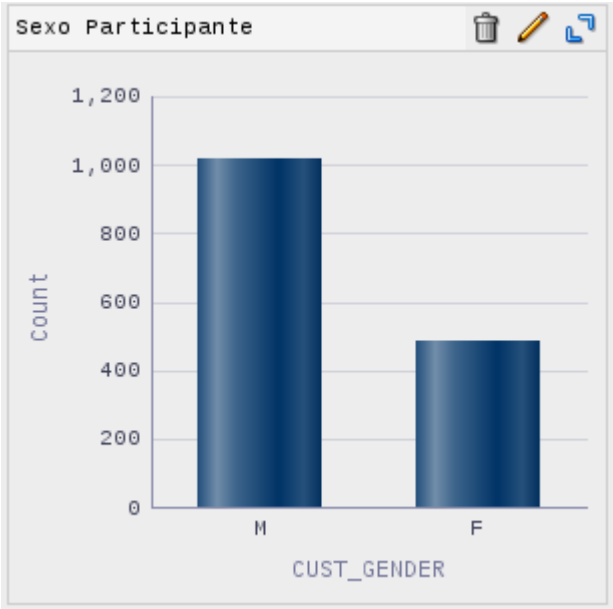
Utilizando o *software* de mineração de dados da Oracle, o *Oracle Data Miner,* conforme tabela 2 da página 26, verificamos a base de dados utilizada em uma campanha anterior com características dos participantes. Dentro deste conjunto de dados, existe uma coluna que contém a idade das pessoas que participaram da campanha. Do total de pessoas, algumas adquiriram o cartão fidelidade, enquanto outras não, e tal informação é encontrada na 2ª coluna da tabela 4 da página 35, denominada *AFFINITY CARD*, onde 1 é para quem adquiriu e 0 para quem não adquiriu, e este grupo de pessoas está distribuído através de idade. Na tabela 4, a 3ª coluna, *AGE,* identifica a idade do participante. Utilizando um histograma, podemos ver qual é a distribuição de idades, conforme figura 5 abaixo:

Figura 5 - Histograma de idade de incorporação ao programa de fidelidade de uma empresa fictícia

Fonte: *Oracle Data Miner,* elaboração própria

A maioria das pessoas que participaram da campanha está entre 24 e 53 anos de idade, e a maioria das pessoas se encontra na faixa dos 31 aos 39 anos de idade, com aproximadamente 300 registros. Outra característica que podemos observar é como os dados estão distribuídos com base no gênero dos participantes.

Figura 6 - Histograma de gênero de incorporação ao programa de fidelidade de uma empresa fictícia



Fonte: *Oracle Data Miner,* elaboração própria

Esta informação é encontrada na 4ª coluna da tabela 5, página 35, denominada *CUST_GENDER,* onde “M” é para masculino e “F” para feminino. Observa-se que o sexo masculino foi muito mais presente durante a campanha realizada, com aproximadamente 1000 registros, enquanto para o sexo feminino obtém-se aproximadamente 500 registros.

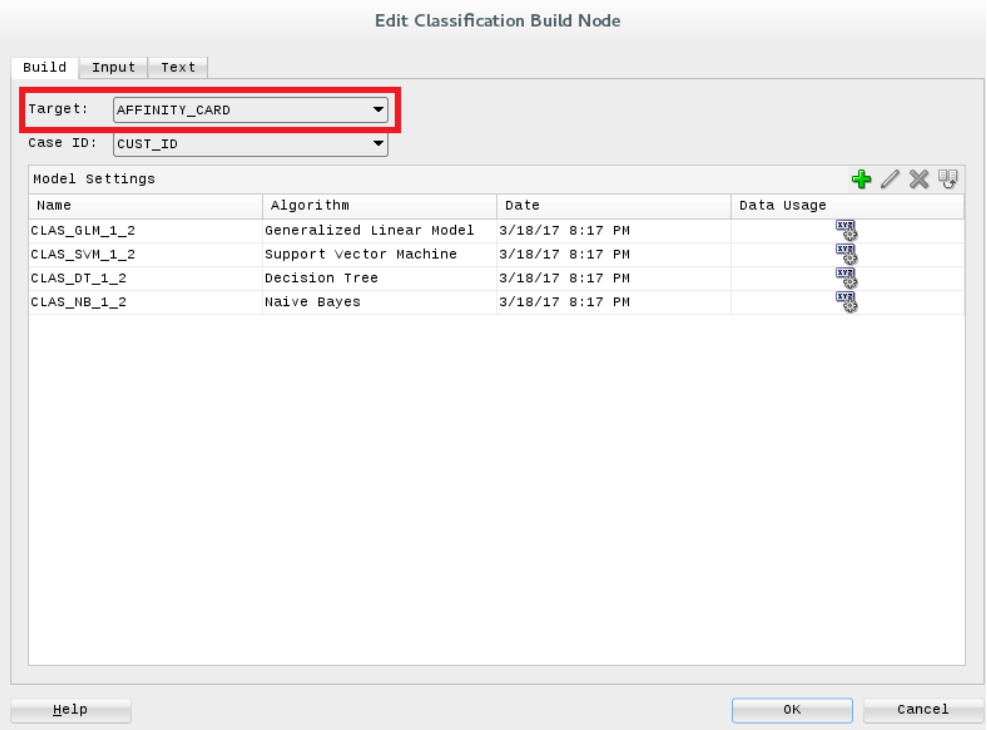
O próximo passo será a criação do modelo de classificação, mas para isso será necessário dividir a base de dados em parte para dados de treino e outra para dados de teste.

Dados de treinamento: é a porcentagem de dados (subconjunto de dados) definida previamente que será utilizada para treinar o algoritmo e criar o modelo. Nesta etapa, utilizam-se os dados históricos em que há a informação se o cliente adquiriu ou não cartão fidelidade, portanto o algoritmo está utilizando esta informação e relacionando com as outras informações dos clientes, representadas pelas colunas da tabela 4 na página 35, para que quando for apresentado a uma nova base de dados, o algoritmo possa prever se o cliente irá ou não adquirir cartão fidelidade baseado nas mesmas características utilizadas na base de dados históricos. Nota-se que para isso, a nova base de dados precisa ter exatamente as mesmas informações contidas na base histórica.

Dados de teste: uma vez que o modelo esteja criado, utiliza-se a porcentagem dados históricos que não foi utilizada pra os dados de treinamento (outro subconjunto), ou seja, os dados de teste. Entretanto, são os dados que o algoritmo não viu ao criar o modelo. O objetivo desta fase é simular um ambiente real, ou seja, serão apresentados dados históricos, mas dados que o modelo ainda não viu. Porém, como já se conhecem os resultados neste conjunto de dados de teste, é possível avaliar a precisão do modelo.

Utilizando a base de dados históricos, definimos a variável alvo que queremos prever, neste caso a coluna de *affinity card,* ou os cartões fidelidade (figura 7).

Figura 7 - Variável alvo e identificador único do modelo de classificação



Fonte: *Oracle Data Miner,* elaboração própria

Esta é a coluna na tabela que indica se o cliente possui ou não possui cartão fidelidade, 1 ou 0, respectivamente. Portanto, é esta a variável que queremos prever para novos clientes em uma nova base de dados, por isso é selecionada a coluna no campo *Target* da ferramenta.

O *Case_ID* é a identificação única para cada registro, podendo ser utilizado o código do cliente, CPF, ou qualquer informação que esteja

disponível que identifique cada pessoa como única. Neste caso, utilizaremos o *cust_id*¸ um valor sequencial gerado automaticamente pela base de dados para cada registro histórico novo cadastrado.

Tabela 5 - Variáveis preditoras

Case ID | Variável alvo (target) | Variáveis preditoras

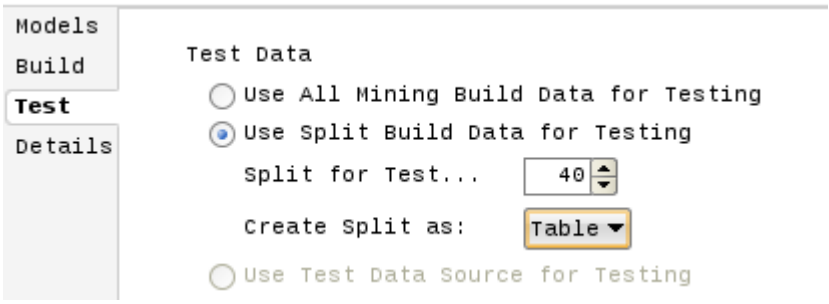
	CUST_ID	AFFINITY_CARD	AGE	CUST_GENDER	EDUCATION	CUST_INCOME_LEVEL	COUNTRY_NAME
1	101,501.0000	0.0000	41.0000	F	Masters	J: 190,000 - 249,999	United States of
2	101,502.0000	0.0000	27.0000	M	Bach.	I: 170,000 - 189,999	United States of
3	101,503.0000	0.0000	20.0000	F	HS-grad	H: 150,000 - 169,999	United States of

Fonte: *Oracle Data Miner*

As demais colunas da tabela, conforme a tabela 5, acima são as variáveis preditoras que irão explicar o comportamento da variável dependente (*affinity_card)*.

Após definir a variável alvo, dividi-se a base de dados para as fases de treinamento e teste, conforme figura 8 da página 41. Dividiremos a base utilizada entre 60% dos dados para treinamento, que serão os dados utilizados pelo algoritmo para construir o modelo, e 40% para teste, que serão os dados que o modelo nunca viu para testar a sua assertividade. Não se usa todos os dados para treinamento para que seja evitado o problema de sobre-ajuste, no qual o modelo pode aprender demais e memorizar resultados. Neste caso, ao utilizar o modelo em uma nova base de dados, o modelo buscaria exatamente as mesmas características dos indivíduos para que o resultado fosse positivo, ou seja, exatamente as mesmas informações de cada coluna, conforme a tabela 4 da página 35. O objetivo é simular um ambiente real onde se conhecem os resultados que o modelo ainda não conhece, sendo possível avaliar a acurácia do mesmo.

Figura 8 - Divisão de dados de treinamento e dados de teste



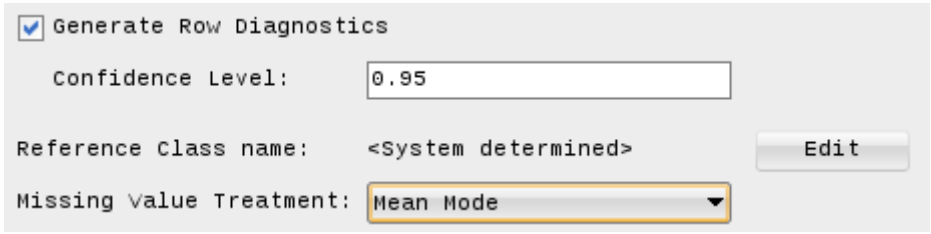
Fonte: *Oracle Data Miner*

Para testar o modelo, primeiramente definem-se as propriedades dos algoritmos que serão utilizados[1].

- GLM: modelo linear generalizado. Este algoritmo trabalha com o tipo de regressão logística, que trabalha com valores categóricos ao invés de numéricos. Primeiramente fazemos um diagnóstico das linhas para verificamos quais possuem valores discrepantes. Para cada variável há um valor médio, desta forma definimos o intervalo de confiança de 95% para eliminarmos os valores discrepantes em relação à média. Também definimos como os campos com valores nulos serão tratados, pois são dados que podem gerar problemas na construção do modelo preditivo. Neste caso, define-se que os dados ausentes serão automaticamente preenchidos com a média dos dados da mesma coluna (variável). A outra opção seria excluir a linha, porém menos dados significa menos observações e consequentemente menos precisão no modelo. A figura 9, abaixo, demonstra o nível de confiança definido e o tratamento de valores nulos, como serão preenchidos, neste caso com a média.

[1] O objetivo deste trabalho consiste em explorar os benefícios do *Big Data* de forma ampla, utilizando-se de algumas técnicas estatísticas. Deste modo, não será aprofundado tecnicamente os métodos estatísticos do qual o *Big Data* pode usufruir para análise de dados, apenas serão apresentadas as suas características gerais.

Figura 9 - Nível de confiança do GLM

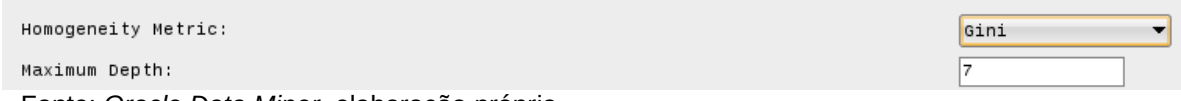


Fonte: *Oracle Data Miner,* elaboração própria.

- SVM: máquinas de vetores de suporte;
    - SVM é um classificador binário e não probabilístico, no qual a partir de um conjunto de dados de entrada, prediz para cada uma das entradas qual de duas possíveis classes a entrada faz parte. Utilizando-se um conjunto de dados de treinamento, cada entrada é marcada como pertencente a uma de duas categorias, e então o algoritmo do SVM faz a representação de novos dados como pontos no espaço que são preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados.
- DT: árvore de decisão;

No algoritmo de Árvore de Decisão, a figura 10 abaixo mostra definição da métrica de homogeneidade como índice de Gini (*homogeneity metric)*, cuja técnica define como os dados serão espalhados pelos “galhos” da árvore. Em seguida, definimos a quantidade de níveis que a árvore terá (*maximum depth)*, totalizando sete níveis. Quanto maior o número de níveis, maior a precisão do modelo, porém mais demorado será para criá-lo. Dependendo da quantidade de dados que está sendo trabalhada, o algoritmo pode levar semanas para ser calculado.

Figura 10 - Índice de Homogeneidade

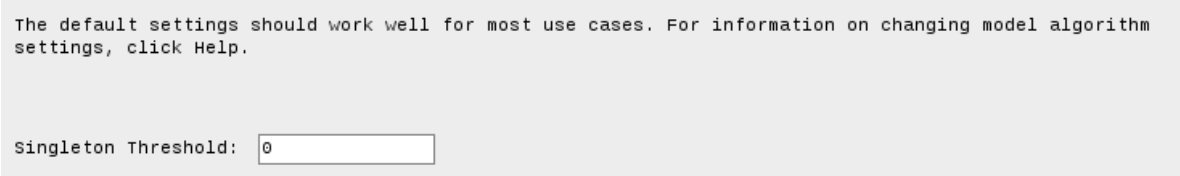


Fonte: *Oracle Data Miner,* elaboração própria.

- NB: Naive Bayes;

No algoritmo de Bayes, definimos o número mínimo de ocorrências de um determinado registro de modo que ele seja ou não incluído no modelo preditivo como uma variável. Por exemplo: se determinada variável aparece apenas duas vezes, ignorar a mesma durante o processo. Neste caso, o *threshold* (limiar) seria igual a 2. No modelo apresentado, definimos como zero, conforme a figura 13 abaixo.

Figura 11 - Limiar do Naive Bayes



Fonte: *Oracle Data Miner,* elaboração própria.

O algoritmo Naive Bayes baseia-se em probabilidades condicionais. Ele usa o Teorema de Bayes, uma fórmula que calcula uma probabilidade contando a freqüência de valores e combinações de valores nos dados históricos.

O teorema de Bayes encontra a probabilidade de ocorrência de um evento, dada a probabilidade de outro evento que já ocorreu. Se B representa o evento dependente e A representa o evento anterior, o teorema de Bayes pode ser indicado da seguinte forma:

Figura 12 - Probabilidade condicional do Teorema de Bayes

$$P(A_j|B) = \frac{P(B \mid A_j) \times P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i) \times P(A_i)}$$

Fonte: Sartoris (2003).

Para calcular a probabilidade de B dado A, o algoritmo conta o número de casos em que A e B ocorrem juntos e o divide pelo número de casos em que A ocorre sozinho.

Para fins de ilustração, a Figura 12 acima mostra um evento dependente baseado em um único evento independente. O algoritmo Naive Bayes geralmente deve levar em conta muitos eventos independentes. Nos dados históricos que utilizamos conforme a tabela 4 da página 35, fatores como idade, gênero, nível de escolaridade, renda e localização do cliente podem ser considerados

Naive Bayes faz a suposição de que cada preditor é condicionalmente independente dos outros. Para um determinado valor alvo, a distribuição de cada preditor é independente dos outros preditores. Na prática, essa suposição de independência, mesmo quando violada, não degrada significativamente a precisão preditiva do modelo e faz a diferença entre um algoritmo rápido e computacionalmente viável.

Após a execução do modelo no *Oracle Data Miner,* todos os algoritmos listados utilizaram 40% da base de dados histórica para treinar, em seguida os outros 60% da mesma base foi utilizado para testar cada um dos algoritmos já treinados e revelar qual foi o melhor algoritmo para este modelo de classificação visando o seu poder de predição.

Inicia-se a comparação dos modelos com base no número de predições corretas. A tabela 6 da página 45 demonstra as seguintes informações: a primeira coluna (*models)* é o algoritmo que foi treinado e testado; a segunda coluna (*correct predictions %)* mostra a porcentagem de acertos do algoritmo com base nos 60% da tabela de dados que foi utilizada para teste; a terceira coluna (*correct predictions count)* é o número de predições corretas; a quarta coluna *(total case count)* é a quantidade de registros totais que representam os 60% da tabela. Ou seja, utilizando o primeiro algoritmo como exemplo, CLAS_NB_1_2, o Naive Bayes, dividindo-se o número de predições corretas pelo total de registros, obtém-se: (499/574)*100 = 86,9338, conforme a segunda coluna demonstra.

Tabela 6 - Precisão dos algoritmos de classificação

Models	Correct Predictions %	Correct Predictions Count	Total Case Count	Total Cost
CLAS_NB_1_2	86.9338	499	574	0.00000000
CLAS_GLM_1_2	76.4808	439	574	0.00000000
CLAS_SVM_1_2	86.9338	499	574	0.00000000
CLAS_DT_1_2	70.2091	403	574	267.74949464

Fonte: *Oracle Data Miner,* elaboração própria.

A quinta coluna da tabela 6 acima, *total cost,* refere-se ao custo, que é analisado a partir de uma matriz de custos e utilizando somente no algoritmo de Árvore de Decisão. Nesta matriz de custo, comparam-se os valores observados no treinamento com os valores previstos no momento do teste Por exemplo, se um modelo classifica um cliente com baixo crédito como de baixo risco, esse erro é caro. A matriz de custos pode tendenciar o modelo a evitar esse tipo de erro. A Matriz de Custo é, basicamente, a combinação de valores observados e valores previstos, onde se podem comparar valores e gerar taxas de acerto, que nos dá os verdadeiro-positivos e falso-positivos, ou seja, aonde deveria ter sido zero (0) e o modelo previu 1, isto é um erro e caracteriza-se como falso-positivo.

Tabela 7 - Matriz de Custo

	0	1
0	256	156
1	15	147
Total	271	303
Correct %	94.4649	48.5149
Cost	63.7156	204.0339

Fonte: *Oracle Data Miner,* elaboração própria.

Acima, a tabela 7 ilustra a matriz de custo gerada pelo algoritmo de Árvore de Decisão do modelo de classificação que foi executado. As colunas representam os valores previstos no momento do teste do modelo, com 60% do conjunto de dados, e as linhas representam os valores reais, os observados no momento do treinamento do modelo com 40% do conjunto de dados. Este conjunto de dados é, ainda, o da Figura 3 da página 32, onde há uma série de características dos clientes e uma das colunas era se o cliente havia ou não adquirido o cartão fidelidade *(affinity card).* Estes dados foram apresentados aos algoritmos que aprenderam os relacionamentos e fizeram previsões, ou seja, dado o conjunto de características do cliente, este cliente deveria ou não adquirir o cartão fidelidade.

Na tabela 7, página 45, os dados históricos utilizados para treinamento estão na linha, onde o valor zero (0) representa os clientes que não adquiriram cartão fidelidade e o valor 1, o cliente que adquiriu o cartão fidelidade. Nesta matriz de

custo, comparam-se os valores observados no treinamento com os valores previstos no momento do teste. O valor zero (0) na primeira coluna indica o valor previsto, ou seja, o modelo acertou 256 vezes para clientes que não adquiriram cartões fidelidade. Olhando para a segunda coluna, em 156 vezes o algoritmo previu o valor 1, quando na verdade deveria ter sido zero (0). Seguindo esta mesma lógica, na segunda linha, onde o valor observado é 1, o modelo previu erroneamente 15 vezes como zero (0), e acertou 147 vezes o valor previsto para 1 onde foi observado este mesmo valor.

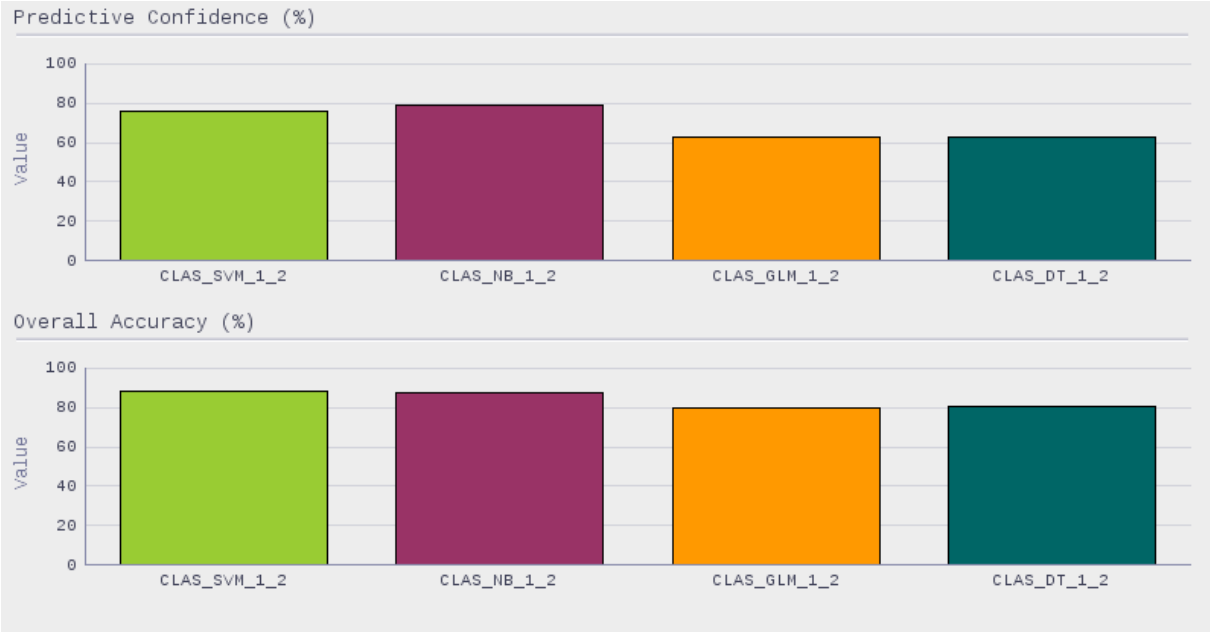
Em seguida, analisa-se a precisão geral, e o nível de confiança dos algoritmos do modelo de classificação.

A precisão geral (*overall accuracy)* refere-se à porcentagem de previsões corretas feitas pelo modelo quando comparado com as classificações reais nos dados do teste.

O nível de confiança *(predictive confidence)* do *Oracle Data Miner* refere-se de forma geral qual é o melhor modelo que se adequa a determinado problema e determinadas características de dados utilizados para criação do modelo preditivo.

A figura 17 ilustra de forma gráfica o nível de confiança e a precisão geral de cada algoritmo utilizado. Abaixo do gráfico, as mesmas informações são apresentadas em tabela com valores em %.

Figura 13 - Algoritmos de Classificação - Confiança e Precisão Geral



Name	Predictive Conf...	Algorithm	Creation Date
CLAS_DT_1_2	62.2660	Decision Tree	3/20/17 7:26 PM
CLAS_GLM_1_2	62.5565	Generalized Lin...	3/20/17 7:26 PM
CLAS_NB_1_2	78.2741	Naive Bayes	3/20/17 7:26 PM
CLAS_SVM_1_2	75.6138	Support Vector ...	3/20/17 7:26 PM

Name	Average Accuracy %	Algorithm	Creation Date
CLAS_DT_1_2	81.1330	Decision Tree	3/20/17 7:26 PM
CLAS_GLM_1_2	81.2782	Generalized Lin...	3/20/17 7:26 PM
CLAS_NB_1_2	89.1371	Naive Bayes	3/20/17 7:26 PM
CLAS_SVM_1_2	87.8069	Support Vector ...	3/20/17 7:26 PM

Fonte: *Oracle Data Miner,* elaboração própria.

Nota-se que os modelos SVM e Naive Bayes tiveram os melhores desempenhos, com 89,1% e 87,8% de acurácia, respectivamente, e com níveis de confiança de 75,6% no SVM e 78,2% no Naive Bayes. Ou seja, para este conjunto de dados, conforme as características utilizadas dos clientes com base nos dados históricos representados pela figura 3 da página 21, o Naive Bayes apresenta a melhor opção como modelo preditivo.

A última etapa é aplicar o modelo a conjuntos de dados totalmente novos onde se quer prever quais clientes irão adquirir cartão fidelidade. A aplicação do modelo nesta etapa utilizará apenas o algoritmo Naive Bayes, pois com base nos

resultados observados anteriormente, é o algoritmo mais confiável a ser utilizado a fim de adquirir melhores previsões.

A nova base de dados deve possuir a mesma estrutura da base histórica utilizada para treino e teste dos algoritmos, ou seja, deverá ter as mesmas colunas como variáveis preditoras e a mesma coluna como *case_id,* conforme a tabela 5 da página 40. A variável alvo *(target),* para este caso, será a coluna gerada pelo *Oracle Data Miner* indicando em cada linha quais dos novos clientes irão adquirir o cartão fidelidade e quais não irão, mostrando também todas as demais características destes clientes, conforme a tabela 8 abaixo.

Tabela 8 - Previsão de Naive Bayes

	AGE	COUNTRY_NAME	EDUCATION	CLAS_NB_1_2_PRED	CLAS_NB_1_2_PROB
1	62.0000	United States of America	< Bach.	0.0000	1.0
2	41.0000	United States of America	Bach.	0.0000	1.0
3	34.0000	United States of America	< Bach.	0.0000	1.0
4	50.0000	United States of America	< Bach.	0.0000	0.7567447423934937
5	46.0000	United States of America	Assoc-A	1.0000	0.9994199872016907
6	20.0000	United States of America	< Bach.	0.0000	0.9999959468841553
7	40.0000	United States of America	HS-grad	0.0000	0.9284946918487549
8	41.0000	United States of America	< Bach.	0.0000	1.0
9	29.0000	United States of America	Bach.	1.0000	1.0
10	28.0000	United States of America	HS-grad	0.0000	1.0
11	31.0000	Brazil	9th	0.0000	1.0
12	35.0000	Singapore	PhD	1.0000	0.9997619986534119
13	42.0000	United States of America	HS-grad	1.0000	0.9665891528129578
14	49.0000	United States of America	HS-grad	0.0000	1.0
15	44.0000	United Kingdom	< Bach.	0.0000	1.0
16	34.0000	United States of America	HS-grad	0.0000	0.9650834202766418
17	68.0000	United States of America	< Bach.	1.0000	0.8164726495742798
18	27.0000	United States of America	< Bach.	0.0000	1.0

Fonte: *Oracle Data Miner,* elaboração própria.

O resultado da previsão com novos dados baseados no modelo preditivo de Naive Bayes mostra, conforme a figura acima, a idade do cliente, o país em que reside, seu nível de educação, se vai ou não adquirir o cartão fidelidade, sendo 1 a opção positiva para adquirir e 0, negativo para adquirir, e por último, qual a porcentagem de chance de acontecer o resultado previsto, na coluna *CLAS_NB_1_2_PROB.* Com estas informações, é possível determinar as principais características que fazem um cliente adquirir um cartão fidelidade, e qual é o público que não adquiriu, oferecendo uma oportunidade à empresa de criar um método para

fidelizar estes clientes com base em suas características pessoais, aumentando, assim, a receita da empresa como um todo.

6. ANALISANDO DADOS EM TEMPO REAL

Este estudo tem como objetivo demonstrar tecnicamente como pode ser possível identificar e prever possíveis casos do vírus Zika a partir de dados que estão sendo coletados em tempo real da rede social Twitter, e alinhar esta informação com o nível de renda da região onde há maior número de citações do vírus em questão com o intuito de identificar possíveis locais com o mesmo perfil onde possa ocorrer o mesmo problema mas que seja possível agir de forma pró ativa e impedir a proliferação da doença. Para isso, usaremos duas ferramentas de análise: *Oracle Big Data Discovery* e *Apache Flume*.

O *Oracle Big Data Discovery* é uma ferramenta da empresa Oracle visualmente intuitiva que permite transformar e visualizar grandes volumes de dados em *insights* a partir de gráficos e mapas. Esta será a ferramenta que usaremos para visualizar a localização das pessoas no mapa e a faixa de renda que se encontram.

O *Apache Flume* é uma ferramenta gratuita, de código aberto para coleta, agregação e movimentação eficientes de grandes quantidades de dados de *logs*, como *tweets*. Possui uma arquitetura simples e flexível baseada em fluxos de dados de transmissão. Esta será a ferramenta que será utilizada para capturar os dados do Twitter.

O projeto ocorrerá em três etapas:

- Coletar os dados da internet a respeito do vírus Zika, especificamente da rede social Twitter, alinhado à localização das pessoas que postam a mensagem para identificar a sua localização no momento da postagem;
- Utilizar uma base de dados do setor censitário proveniente do IBGE, com o intuito de identificar as faixas de rendas em salários mínimos de cada

perímetro do Paraná definido pelo Instituto e plotar esta informação em um mapa segregado em cores que representam as faixas de renda que serão definidas posteriormente neste trabalho;

- Mesclar as duas informações, a localização das pessoas que estão postando sobre o vírus na internet (fonte Twitter) com a localização de sua faixa de renda (fonte IBGE), e identificar a possibilidade de prever um surto com base no número de indivíduos postando na mesma região e se há alguma relação com a faixa de renda do setor.

6.1 Base de dados 1: Coletando dados do *Twitter*

O primeiro passo será monitorar os *tweets* sendo feitos na internet a partir de *hashtags* que serão configuradas para pegar os *tweets*. *Hashtag "*é um composto de palavras-chave, ou de uma única palavra, que é precedido pelo símbolo cerquilha (#). *Tags* significam etiquetas e referem-se a palavras relevantes, que associadas ao símbolo # se tornam *hashtags* que são amplamente utilizadas nas redes sociais, em especial no Twitter, onde a adesão delas as tornou tão popular. Esse tipo de marcação, utilizada nas redes sociais e em outros meios, serve para associar uma informação a um tópico ou discussão. Geralmente essas *hashtags* tornam-se links indexáveis pelos mecanismos de busca. Isso permite que os demais usuários possam clicar nelas ou procurá-las e visualizarem todas as informações, imagens, vídeos, entre outros relacionados a elas." (Canaltech, 2017).

Para este processo, utiliza-se um arquivo padrão da ferramenta *Flume* em formato texto onde constam os parâmetros para definições das *hashtags* a serem monitoradas, conforme a figura 14 abaixo.

Figura 14 - Arquivo de configuração de coleta de dados do *Twitter*

```
# The configuration file needs to define the sources,
# the channels and the sinks.
# Sources, channels and sinks are defined per agent,
# in this case called 'TwitterAgent'

TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = <required>
TwitterAgent.sources.Twitter.consumerSecret = <required>
TwitterAgent.sources.Twitter.accessToken = <required>
TwitterAgent.sources.Twitter.accessTokenSecret = <required>
TwitterAgent.sources.Twitter.keywords = zika, zikavirus
TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://hadoop1:8020/user/flume/tweets/%Y/%m/%d/%H/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Fonte: Elaboração própria

A linha realçada em azul, na figura 14 acima, é a linha onde se define a *hashtag* a ser monitorada, e neste caso definimos duas palavras-chaves: “zika” e “zikavírus”. Isso significa que qualquer texto publicado no *Twitter* que contenha alguma destas duas palavras será gravado em formato de arquivo *log* dentro de uma base de dados. Os demais campos deste documento são parâmetros técnicos que serão lidos pelo programa e são normalmente utilizados em seu formato padrão, sem nenhuma modificação.

O próximo passo é criar uma tabela na base de dados com as colunas que desejamos ver as informações.

Figura 15 - Criação da tabela de armazenamento de dados do *Twitter*

```
CREATE EXTERNAL TABLE tweets_zika (
id BIGINT,
created_at STRING,
source STRING,
favorited BOOLEAN,
retweet_count INT,
retweeted_status STRUCT<
text:STRING,
user:STRUCT<screen_name:STRING,name:STRING>>,
entities STRUCT<
urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>,
text STRING,
user STRUCT<
screen_name:STRING,
name:STRING,
friends_count:INT,
followers_count:INT,
statuses_count:INT,
verified:BOOLEAN,
utc_offset:INT,
time_zone:STRING>,
in_reply_to_screen_name STRING
)
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
LOCATION '/user/flume/tweets'
```

Fonte: Elaboração própria

O código ilustrado na figura 15 acima representa a criação de uma tabela na base de dados denominada ***tweets_zika***. Esta tabela possui como principais informações relevantes as seguintes colunas:

Created_at: data e hora que o *tweet* foi postado;

Source: qual foi a fonte geradora do dado (iphone, PC, android);

Text: o texto escrito pelo usuário;

User: o nome do usuário que escreveu o texto;

Friends_count: a quantidade de amigos que o usuário que escreveu o texto possui;

Followers_count: a quantidade de seguidores que o usuário que escreveu o texto possui;

***Time_zone:** nome da cidade que a pessoa estava quando digitou o texto.

Time_zone: é possível pegar a informação da localização da pessoa que *tweetou* com base no GPS do dispositivo, em caso de celular, ou com o número do IP da rede, no caso de computadores pessoais.

Em seguida, é feito o seguinte comando no sistema que se encontra o agente do *Flume* para iniciar a coleta de dados:

Figura 16 - Execução do agente de captura de *tweets*

```
oracle@bigdatalite flume]$ flume-ng agent --conf ./ -f flume.conf -Dflume.root.logger=DEBUG,console -n TwitterAgent
```

Fonte: Elaboração própria

Este comando, conforme a figura 16 acima inicia o processo de monitoramento do *Flume* no *Twitter* com base nas configurações definidas conforme o modelo de arquivo da figura 21 na página 59.

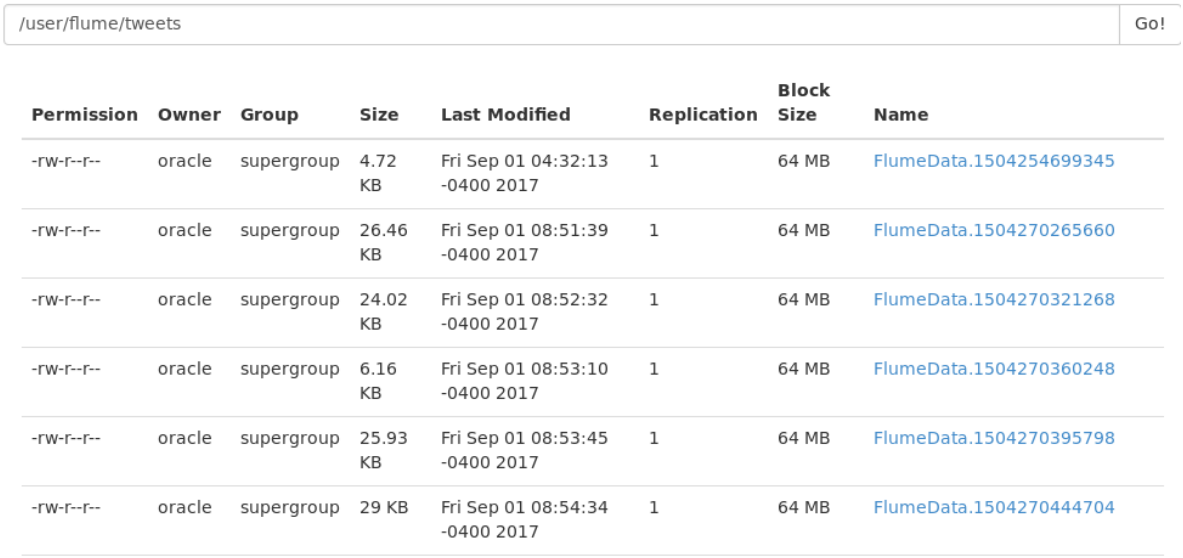
Figura 17 - Representação da captura de um *tweet* em tempo real

```
2017-09-17 15:34:36,400 (hdfs-HDFS-call-runner-0) [DEBUG - org.apache.htrace.core.Tracer$Builder.loadSamplers(Trace
r.java:106)] sampler.classes = ; loaded no samplers
2017-09-17 15:34:36,405 (hdfs-HDFS-call-runner-0) [DEBUG - org.apache.htrace.core.Tracer$Builder.loadSpanReceivers(
Tracer.java:128)] span.receiver.classes = ; loaded no span receivers
2017-09-17 15:34:40,657 (hdfs-HDFS-call-runner-0) [DEBUG - org.apache.flume.sink.hdfs.AbstractHDFSWriter.reflectGet
NumCurrentReplicas(AbstractHDFSWriter.java:199)] Using getNumCurrentReplicas--HDFS-826
2017-09-17 15:34:40,665 (hdfs-HDFS-call-runner-0) [DEBUG - org.apache.flume.sink.hdfs.AbstractHDFSWriter.reflectGet
DefaultReplication(AbstractHDFSWriter.java:227)] Using FileSystem.getDefaultReplication(Path) from HADOOP-8014
2017-09-17 15:35:10,668 (hdfs-HDFS-roll-timer-0) [DEBUG - org.apache.flume.sink.hdfs.BucketWriter$2.call(BucketWrit
er.java:276)] Rolling file (hdfs://bigdatalite.localdomain:8020/user/flume/tweets/FlumeData.1505676875037.tmp): Rol
l scheduled after 30 sec elapsed.
2017-09-17 15:35:10,672 (hdfs-HDFS-roll-timer-0) [INFO - org.apache.flume.sink.hdfs.BucketWriter.close(BucketWriter
.java:363)] Closing hdfs://bigdatalite.localdomain:8020/user/flume/tweets/FlumeData.1505676875037.tmp
2017-09-17 15:35:10,751 (hdfs-HDFS-call-runner-4) [INFO - org.apache.flume.sink.hdfs.BucketWriter$8.call(BucketWrit
er.java:629)] Renaming hdfs://bigdatalite.localdomain:8020/user/flume/tweets/FlumeData.1505676875037.tmp to hdfs://
bigdatalite.localdomain:8020/user/flume/tweets/FlumeData.1505676875037
2017-09-17 15:35:10,770 (hdfs-HDFS-roll-timer-0) [INFO - org.apache.flume.sink.hdfs.HDFSEventSink$1.run(HDFSEventSi
nk.java:394)] Writer callback called.
```

Fonte: Elaboração própria

Após alguns segundos com o *Flume* em execução, é possível ver a seguinte mensagem realçada em amarelo na figura 17 acima. Esta é a representação de que um dado está sendo coletado no formato de *log* e foi gravado na base de dados. Dentro deste dado há as informações que queremos analisar conforme a tabela de armazenamento de dados criada na figura 23 da página 59. É neste momento que se captura as informações em tempo real e é possível analisá-las conforme vão sendo coletadas para a base de dados.

Figura 18 - Base de dados com *logs* dos *tweets*



/user/flume/tweets Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	oracle	supergroup	4.72 KB	Fri Sep 01 04:32:13 -0400 2017	1	64 MB	FlumeData.1504254699345
-rw-r--r--	oracle	supergroup	26.46 KB	Fri Sep 01 08:51:39 -0400 2017	1	64 MB	FlumeData.1504270265660
-rw-r--r--	oracle	supergroup	24.02 KB	Fri Sep 01 08:52:32 -0400 2017	1	64 MB	FlumeData.1504270321268
-rw-r--r--	oracle	supergroup	6.16 KB	Fri Sep 01 08:53:10 -0400 2017	1	64 MB	FlumeData.1504270360248
-rw-r--r--	oracle	supergroup	25.93 KB	Fri Sep 01 08:53:45 -0400 2017	1	64 MB	FlumeData.1504270395798
-rw-r--r--	oracle	supergroup	29 KB	Fri Sep 01 08:54:34 -0400 2017	1	64 MB	FlumeData.1504270444704

Fonte: Elaboração própria

Nota-se que, após algum tempo, já se possuem vários dados coletados no formato **FlumeData.###**, conforme figura 18 acima. Também temos a informação de quando o arquivo foi criado, ou seja, coletado do Twitter, na 4ª coluna denominada **Last Modified.** O processo de coleta em tempo real não cessa enquanto não for executado um comando para o mesmo parar. Durante este tempo, toda informação já é disponibilizada na base de dados conforme o agente do *Flume* captura o dado.

Com os dados já inseridos em nossa base de dados, podemos selecionar a tabela *tweets2*, criada anteriormente conforme a figura 15 da página 54 para verificar as informações contidas nele.

Tabela 9 - Tabela com informações dos dados coletados do *twitter*

tweets_zika.created_	tweets_zika.source	tweets_zika.text	tweets_zika.user	tweets_zika.time_zone
Sun Sep 17 19:38:03	">Twitter for Android	@thismorning Thinks it's a pure piss take of an ugly version of s	{"screen_name":"tashaR2017","name":"Tasha Roylance",'	ll,"time_zone":null}
Sun Sep 17 19:45:38	w">Twitter for iPhone	@ocedlc Quel est le dernier tableau qu'à vu lady diana ?	{"screen_name":"AynilBrtn","name":"Gabin","friends_cou	ll,"time_zone":null}
Sun Sep 17 19:51:09	w">Twitter for iPhone	dahora zika kkkk	{"screen_name":"hgosvage","name":"nestot","friends_cou	one":"Mid-Atlantic"}
Sun Sep 17 19:51:43	">Twitter for Android	5 gols fora o show, a zika saiu	{"screen_name":"lipecamargo_","name":"feLIPE","friends_	me_zone":"Brasilia"}
Sun Sep 17 19:51:53	">Twitter for Android	Wesley trouxe a zika dele e do São Paulo. Jogador morto.	{"screen_name":"tarcioomendes","name":"mendes","frien	ll,"time_zone":null}
Sun Sep 17 19:51:55	w">Twitter for iPhone	sai zika	{"screen_name":"santos8tao","name":"fernanda","friends	ll,"time_zone":null}
Sun Sep 17 19:53:17	w">Twitter for iPhone	Respeita a zika	{"screen_name":"eosereia","name":"Lenis ","friends_cou	Time (US & Canada)"}
Sun Sep 17 19:53:26	">Twitter for Android	RT @Sremski_Front: Brate bolje su uskladjeni pijani Zika i Pera	{"screen_name":"mirjanamirka6","name":"Mirjana Malis	ll,"time_zone":null}
Sun Sep 17 19:54:01	">Twitter for Android	RT @marcwscrf_: Deus tira essa zika do Diego pfvr, ta bom já el	{"screen_name":"caliciamaycrf","name":"may","friends_c	Time (US & Canada)"}
Sun Sep 17 19:54:51	">Twitter for Android	Não tô nenhum pouco afim de ir pro Florence, mais o Marcelo é	{"screen_name":"la_loshua","name":"Laura","friends_cou	me_zone":"Brasilia"}
Sun Sep 17 19:55:07	ofollow">Twitter Lite	O batman é zika pq ele bate nos maluco do psol petista fdp	{"screen_name":"rafilhodaputa","name":"Rafiladapreula'	ll,"time_zone":null}
Sun Sep 17 19:55:45	">Twitter for Android	Izinto zika life ke ezo https://t.co/lzOUD3w3Uw	{"screen_name":"Penxenxe","name":"The Villager","friend	me_zone":"Pretoria"}
Sun Sep 17 19:55:52	w">Twitter for iPhone	@joeflore Zika reversa kkkkkkk	{"screen_name":"danielrib9","name":"DANIEL FABULOSO	time_zone":"Hawaii"}
Sun Sep 17 19:55:55	">Twitter for Android	Vamo fazer uma corrente de oração pra tirar a zika ue colocara	{"screen_name":"silvaa_luannaa","name":"Luana the Wit	Time (US & Canada)"}
Sun Sep 17 19:55:55	">Twitter for Android	@fernandokallas Hahaha que comemoração maravilhosa e ape	{"screen_name":"correiacool","name":"douglas silva","fr	one":"Mid-Atlantic"}

Fonte: Elaboração própria

Nota-se, pela tabela 11 acima, que as colunas estão de acordo com as criadas conforme a figura 23 da página 59, com destaque para a 3ª e 5ª coluna, texto e localização, respectivamente. Também é possível notar em alguns textos que a palavra “zika” não está diretamente relacionada com o vírus em si. Este pode ser um problema a ser sanado com a calibração das palavras-chaves, mas para este trabalho que tem como objetivo ilustrar o processo de coleta e análise de dados em tempo real trabalha-se de forma hipotética com as palavras-chave previamente configuradas.

Com os dados coletados, o próximo passo é configurar o mapa e carregar estes dados do Twitter conforme a sua localização para identificar de onde estavam sendo *tweetados* e qual a faixa de renda da região.

6.2 Base de dados 2: Setor censitário IBGE

A ferramenta da empresa Oracle, *Big Data Discovery*, já possui um mapa padrão, porém, sem a separação dos setores conforme a renda das residências em um determinado perímetro. Para construirmos este mapa, utilizaremos como base o site do IBGE que disponibiliza uma Base de Informações por Setor Censitário. “Os dados deste arquivo, por setor censitário, compreendem características dos domicílios particulares e das pessoas que foram investigadas para a totalidade da população e são denominados, por convenção, resultados do universo. Estes dados

foram obtidos reunindo informações captadas por meio da investigação das características dos domicílios e das pessoas, que são comuns aos dois tipos de questionários utilizados para o levantamento do Censo Demográfico 2010 e que são:

· **Questionário Básico** - aplicado em todas as unidades domiciliares, exceto naquelas selecionadas para a amostra, e que contém a investigação das características do domicílio e dos moradores; e

· **Questionário da Amostra** - aplicado em todas as unidades domiciliares selecionadas para a amostra. Além da investigação contida no Questionário Básico, abrange outras características do domicílio e pesquisa importantes informações sociais, econômicas e demográficas dos seus moradores. O setor censitário é a menor unidade territorial, formada por área contínua, integralmente contida em área urbana ou rural, com dimensão adequada à operação de pesquisas e cujo conjunto esgota a totalidade do Território Nacional, o que permite assegurar a plena cobertura do País.” (IBGE, 2010).

A planilha *DomicílioRenda_UF.xls*, na tabela 12 abaixo, fornece informação sobre os rendimentos dos domicílios, pessoas e responsáveis.

Tabela 10 - Variáveis de renda do IBGE

Cod_setor	Situacao_setor	V001	V002	V003	V004	V005	V006	V007	V008	V009	V010	V011	V012	V013	V014	Total 9 até 13
410010305000001	1	1	529942	529432	510	1	7	43	117	86	25	12	3	2	3	128
410010305000002	1	0	611161	611161	0	0	9	29	125	83	29	12	9	3	3	136
410010305000003	1	0	446825	446825	0	3	41	105	131	57	4	4	1	2	50	68
410010305000004	1	1	313176	312516	660	1	20	75	126	51	5	1	0	0	12	57
410010305000005	1	2	433367	431706	1661	2	11	65	141	87	9	2	0	0	5	98
410010305000006	8	0	84367	84367	0	2	7	20	33	8	1	0	0	0	4	9
410010305000007	8	0	90502	90502	0	1	8	18	28	6	2	2	1	0	3	11
410010305000008	8	1	147902	147602	300	6	15	32	41	20	4	1	2	0	4	27
410010305000009	8	0	118829	118829	0	1	9	30	25	10	1	2	2	2	7	17
410010305000010	8	0	192305	192305	0	2	14	36	69	36	1	2	0	0	2	39

Fonte: http://www.ibge.gov.br/

Esta planilha possui a variável de identificação do setor censitário (Cod_setor) na 1ª coluna. Cada linha da planilha fornece os dados de um setor censitário e cada coluna corresponde a uma variável, seja o código ou nome de uma subdivisão geográfica, seja o tipo ou situação do setor, seja, ainda, o valor numérico de uma variável de domicílio, responsável ou pessoa, conforme a figura abaixo.

Tabela 11 - Descrição das variáveis do IBGE

NOME DA VARIÁVEL	DESCRIÇÃO DA VARIÁVEL
Cód_setor	Código do setor censitário
Situação	Código de situação do setor censitário (ver planilha Basico_UF.xls)
V001	Total de domicílios particulares improvisados
V002	Total do rendimento nominal mensal dos domicílios particulares
V003	Total do rendimento nominal mensal dos domicílios particulares permanentes
V004	Total do rendimento nominal mensal dos domicílios particulares improvisados
V005	Domicílios particulares com rendimento nominal mensal domiciliar per capita de até 1/8 salário mínimo
V006	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1/8 a 1/4 salário mínimo
V007	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1/4 a 1/2 salário mínimo
V008	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1/2 a 1 salário mínimo
V009	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 1 a 2 salários mínimos
V010	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 2 a 3 salários mínimos
V011	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 3 a 5 salários mínimos
V012	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 5 a 10 salários mínimos
V013	Domicílios particulares com rendimento nominal mensal domiciliar per capita de mais de 10 salários mínimos
V014	Domicílios particulares sem rendimento nominal mensal domiciliar per capita

Fonte: http://www.ibge.gov.br/

Para este projeto, utilizaremos as variáveis: v009, v010, vv011, v012 e v013, conforme figura X, acima.

Além da planilha de dados, foi utilizado um arquivo com informações adicionais relacionadas aos setores censitários e à divisão territorial:

Descrição dos SetoresPR.xls, com informações sobre o ponto inicial e o perímetro de cada setor censitário, além da indicação de áreas ou setores, contidos no perímetro, que não pertencem ao setor censitário, e dos aglomerados rurais contidos no setor, que podem ou não pertencer ao setor.

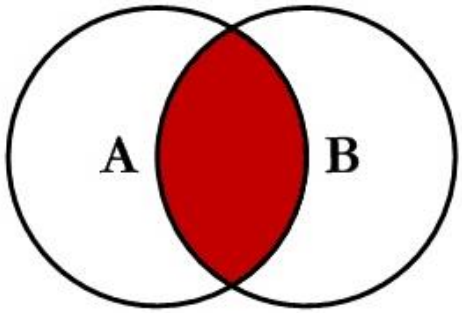
Tabela 12 - Geocódigos do IBGE

Geocodigo	uf	mun	distr	sdist	setor	ponto inicial
410040005000009	41	00400	05	00	0009	RUA DELEGADO THEOLINDO BATISTA DE SIQUEIRA COM RUA IRENE COLODEL DA CRUZ.
410040005000010	41	00400	05	00	0010	RUA IZIDORO DE PAULA COM A RUA TENENTE JOSE TENORIO DE ALBUQUERQUE
410040005000011	41	00400	05	00	0011	RUA I COM PROPRIEDADE DOS HERDEIROS DE GODOFREDO DE SOUZA MACHADO
410040005000012	41	00400	05	00	0012	RUA LAURO BATISTA DE SIQUEIRA COM RUA ANTÔNIO BATISTA SIQUEIRA
410040005000013	41	00400	05	00	0013	RUA LOURENÇO ANGELO BUZATO COM RUA ARI ANTÔNIO BUZATO
410040005000014	41	00400	05	00	0014	RODOVIA DOS MINÉRIOS COM A RUA ETORE BUZATO
410040005000015	41	00400	05	00	0015	RUA SEM DENOMINAÇÃO COM A RUA PEDRO TEIXEIRA ALVES
410040005000016	41	00400	05	00	0016	RUA ANTÔNIO FERRO OU RUA IZIDORO PEDRO BUZATO COM ESTRADA BOICHININGA
410040005000017	41	00400	05	00	0017	ESTRADA BOICHININGA COM RUA ANTÔNIO FERRO
410040005000018	41	00400	05	00	0018	RODOVIA TAMANDARÉ-COLOMBO COM A RUA ANTÔNIO FERRO

Fonte: http://www.ibge.gov.br/

A 1ª coluna da tabela 14 acima, “Geocodigo”, é a coluna utilizada para identificar o polígono que se encontra em determinada variável da planilha *DomicílioRenda_UF.xls* na página 61, quando combinada com a coluna do Cod_setor da planilha *DomicílioRenda_UF.xls,* desta forma temos a informação conjunta dá área em que a residência se encontra e a faixa de renda predominante no perímetro. Toda esta transformação é feita dentro da própria ferramenta, apenas fazendo uma ligação entre as duas tabelas utilizando colunas que possuem a mesma informação. O nome deste processo é dado como *join,* e sua representação conceitual segue na figura 27, abaixo:

Figura 19 - Representação de um *join* entre conjuntos de dados



Fonte: Elaboração própria.

O Inner Join é o método de junção mais conhecido e, como ilustra a Figura 19, retorna os registros que são comuns às duas tabelas.

As informações das variáveis utilizadas (v009, v010, v011, v012 e v013), foram tratadas e formatas condicionalmente da seguinte forma:

v009 – cor amarela (1 a 2 salários mínimos);

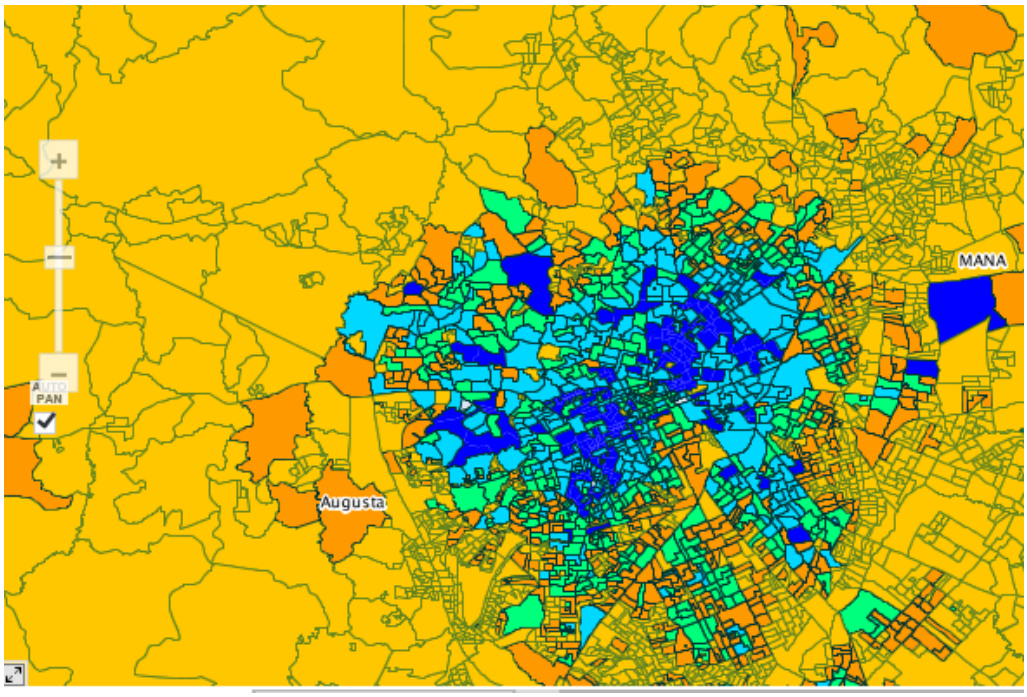
v010 – cor laranja (2 a 3 salários mínimos);

v011 – cor verde (3 a 5 salários mínimos);

v012 – cor azul claro (5 a 10 salários mínimos);

v013 – cor azul escuro (mais de 10 salários mínimos);

Figura 20 - Mapa do setor censitários segregado por renda



Fonte: *Big Data Discovery.* Elaboração própria

1 a 2 salários mínimos
2 a 3 salários mínimos
3 a 5 salários mínimos
5 a 10 salários mínimos
Mais de 10 salários mínimos

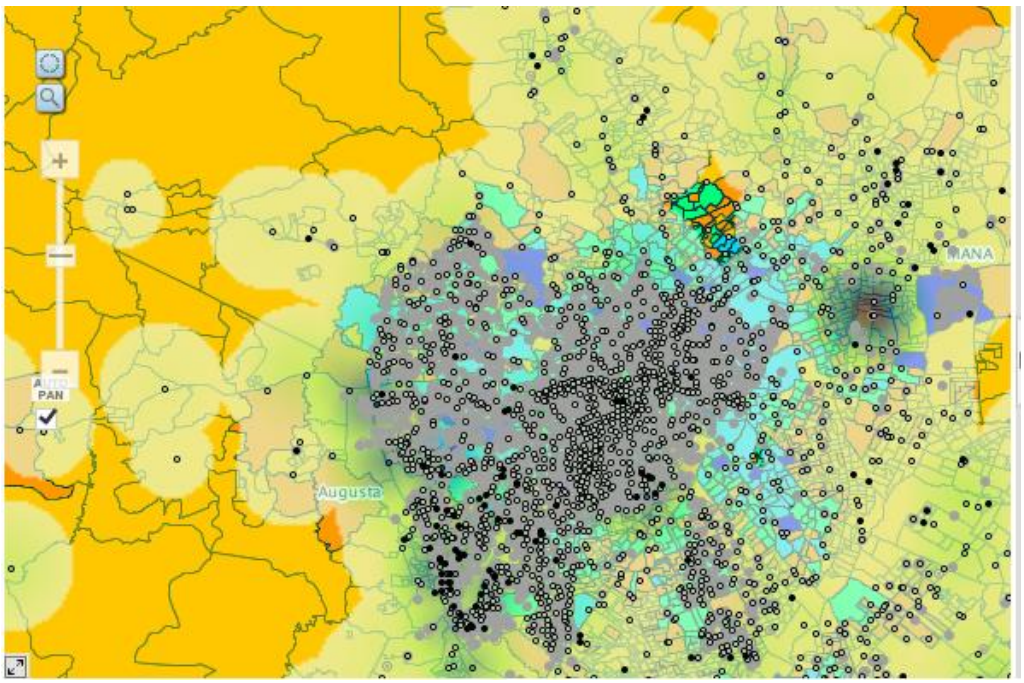
Após estas transformações temos o mapa do Paraná segregado por perímetros que respeitam as regras das variáveis e suas respectivas cores conforme a legenda. A figura 20, acima, demonstra a cidade de Curitiba e alguns de municípios vizinhos.

O próximo passo será incluir a localização de cada pessoa que fez um *tweet* para que seja possível identificar de onde vieram a maioria das informações dentro do Estado do Paraná como um todo.

6.3 Análise de dados do setor censitário e *tweets*

O primeiro passo neste momento é, utilizando a própria ferramenta, transformar o campo de *time_zone* em latitude e longitude para que cada *tweet* tenha um ponto no mapa onde possamos identificar a área a qual pertence. Com este processo feito, adicionamos a informação no tipo de visualização do mapa e temos o seguinte resultado:

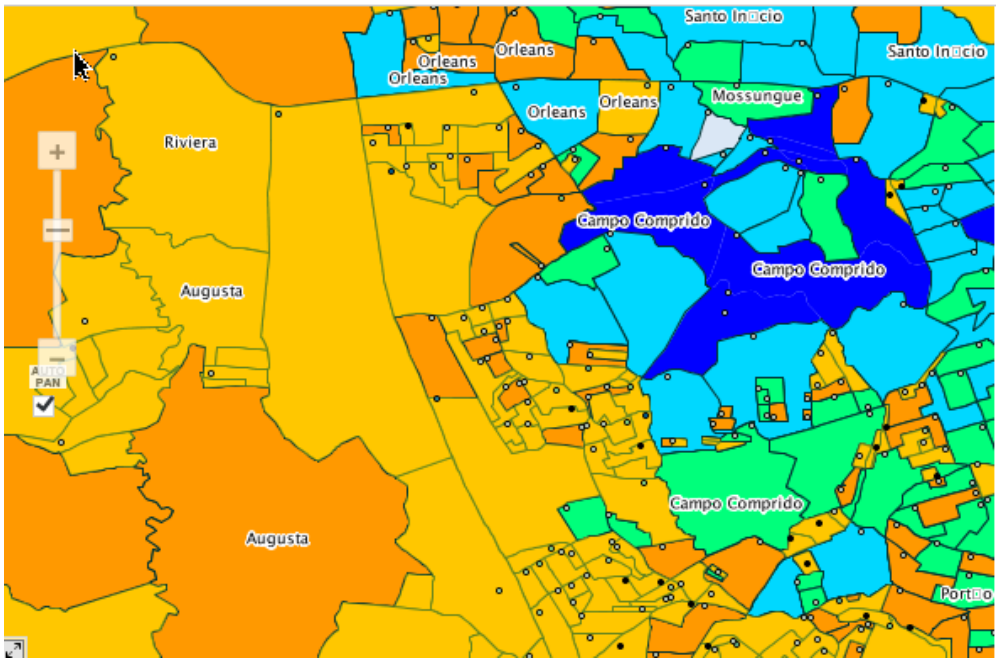
Figura 21 - Mapa do setor censitário e *tweets*



Fonte: *Big Data Discovery.* Elaboração própria

Cada ponto cinza representa a latitude e longitude de uma pessoa que realizou um *tweet* utilizando as palavras-chaves “zika”, “zikavirus”, conforme definido na página 57.

Figura 22 - Mapa do setor censitário e *tweets (zoom)*



Fonte: *Big Data Discovery.* Elaboração própria

1 a 2 salários mínimos
2 a 3 salários mínimos
3 a 5 salários mínimos
5 a 10 salários mínimos
Mais de 10 salários mínimos

Ao aproximar o mapa ainda mais, temos uma melhor noção de onde vem a maior dos *tweets* e qual o nível de renda predominante da área. No caso da imagem acima, pode-se perceber que o bairro Augusta, com predominância das variáveis v009 e v 010, possui uma grande concentração de pessoas que publicaram um *tweet* a respeito do Zika Vírus, ou seja, neste caso, de uma área onde a maioria das residências possuem de 1 a 3 salários mínimos, ao contrário das áreas azuis, onde há poucos pontos e a faixa de renda é de pelo menos 5 salários mínimos.

Este caso de estudo buscou ilustrar como é possível alinhar duas fontes de informações distintas (*Twitter* e IBGE), incluindo a coleta de dados constante e em tempo real, para tentar prevenir a doença em determinada região onde está sendo mais comentada e também repassar a informação de localização e nível de renda da maioria das residências desta região. Esta poderia ser uma informação de grande

valia para o Ministério da Saúde, por exemplo, para identificar áreas com o mesmo perfil censitário e, assim, agir pró-ativamente e evitar que o vírus se prolifere.

CONCLUSÃO

O desenvolvimento deste trabalho possibilitou explorar os conceitos de *Big Data*, levando em conta as características conhecidas como os 5 V's da informação: volume, variedade, velocidade, veracidade e valor, e como o setor privado e o setor público podem fazer uso deste conceito para auxiliá-los na elaboração de políticas e ações públicas e bem como o setor privado na tomada de decisões empresariais.

Além disso, também permitiu uma demonstração mais consistente sobre as etapas do processo, para obter um melhor grau de conhecimento em *Big Data*.

Percebeu-se durante o trabalho que o processo de descoberta de conhecimento em bases de dados passa por uma série de etapas, desde a coleta, a mineração dos dados, a consolidação e extração dos padrões e regras, e por fim a agregação de valor que possibilita uma melhor tomada de decisão.

De um modo geral percebe-se um grande potencial na utilização do *Big Data* como ferramenta auxiliar para elaboração de políticas públicas como por exemplo na área de saúde publica através do uso da informação que está sendo gerada a todo o momento na internet e pode ser utilizada como fonte de dados para prevenção de epidemias, como foi o caso do vírus Zika. Simples menções do estado de saúde do cidadão em redes sociais podem ser valiosas para identificar pontos de atenção.

Nesse sentido, a utilização de recursos digitais adequados permite com que a população mais necessitada seja atendida não só mais rapidamente, mas com a quantidade de recursos suficientes e adequados para atendê-los gerando assim grande assertividade e com um melhor emprego dos recursos públicos.

Com a utilização de duas técnicas para se utilizar do *Big Data*, classificação e regressão, percebemos que há possibilidade de aumentar significativamente a

assertividade e a precisão de previsões ao se utilizar as ferramentas corretas com a grande quantidade de dados disponíveis.

É notável o desenvolvimento da tecnologia e o crescimento do uso da mesma por parte da população através de dispositivos eletrônicos. Com cada vez mais dados sendo gerados, armazenados e utilizados como fonte de informação, percebe-se, através dos exemplos desenvolvidos neste trabalho, que o uso do *Big Data* pode ser promissor para diversas análises e contribuir com melhor embasamento e fundamentos empíricos a ampliar a capacidade de agir e tomar decisões.

Dada a importância do assunto, torna-se necessário um maior aprofundamento do estudo empregando a utilização de mais dados, tornando as análises mais assertivas e fáceis de serem interpretadas, podendo assim economizar tempo entre a identificação do problema, as ações a serem tomadas e os recursos que são necessários para serem concluídas.

REFERÊNCIAS

BARBIERI, C. – **BI – Business Intelligence: Modelagem e Tecnologia**. 2ª ed. Rio de Janeiro: Axcel Books do Brasil Editora, 2001.

BRIETMAN, K. – ***Big Data Overview*. EMC *Summer School on Big Data***. EMC/NCE/UFRJ. Rio de Janeiro. 2013. Disponível em: http://2014.emcbigdataschool.nce.ufrj.br/images/presentations/_Big_Data_Summer_School_Karin.pdf.

CARR, D. ***Giving Viewers What They Want.*** New York Times, 2013. Disponível em: http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html?mcubz=0

CASTELLS, M. **A Sociedade em Rede, Vol. 1.** Editora Paz e Terra, 2000.

DAVENPORT, T. H. – ***Big Data at Work: Dispelling the Myths, Uncovering the Opportunities****. Harvard Business Review Press Books*. 2014.

DAVENPORT, T. H.; PATIL, D.J. – ***Data Scientist: The Sexiest Job of the 21st Century.*** *Harvard Business Review* 90, no. 10, *October*, p.70–76, 2012.

DUTRA, R. G. – **Aplicação de Métodos de Inteligência Artificial em Inteligência dos Negócios.** XXV ENEGEP, Porto Alegre – RS, 2005.

FAYYAD, U.; SHAPIRO, G. P. – ***From Data Mining to knowledge Discovery in databases***. AI. Magazine, 17, Fall 1996.

FLORISSI, P. – ***Big Data*. EMC *Corporation on Big Data***. 2012. Disponível em: https://www.carecorenational.com/healthcaresummit/powerpoints/PatriciaFlorissiPhD.pdf

FROST & SULLIVAN. ***Latin America Big Data and Analytics Market, Forecast to 2022.*** Disponível em: https://www.reportbuyer.com/product/4683703/latin-america-big-data-and-analytics-market-forecast-to-2022.html

GIUDICI, P. – ***Applied Data Mining: statistical methods for business and Industry***. John Wiley & Sons Ltd. 2003.

GOLDSCHMIDTH, R.; PASSOS, E. – ***Data Mining*: Um Guia Pratico.** Rio de Janeiro: Elsevier, 2005, 3ª reimpressão.

GONÇALVES, E. C. – **Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas**. INFOCOMP (UFLA), 2005.

GRILO JÚNIOR, T. F. – **Aplicação de Técnicas de *Data Mining* para Auxiliar no Processo de Fiscalização no Âmbito do Tribunal de Contas do Estado da Paraíba** – UFPB, 2010.

HEATH T. & BIZER C. – ***Linked Data: Evolving the Web into a Global Data Space* (1st edition). *Synthesis Lectures on the Semantic Web*: *Theory and Technology***. Morgan & Claypool Publishers, 2011.

HYUGAN, May. ***Introduction to Big Data***, 2015. Disponível em: https://www.coursera.org/learn/big-data-introduction/.

IGNACIO, S. A. – **Importância da Estatística para o Processo de Conhecimento e Tomada de Decisão**. Revista Paranaense de Desenvolvimento, Curitiba, nº 118 – jan/jun 2010.

KOH, H - ***How Big Data Has Changed Public Policy***, 2017. Disponível em: https://dataflog.com/read/how-big-data-has-changed-public-policy-infographic/1880

MARCHAND, D. A.; PEPPARD, J. – ***Why IT Fumbles Analytics**. Harvard Business Review*, 2013.

MATTOSO, M. – ***Scientific Workflows and Big Data*. EMC *Summer School on Big Data***. EMC/NCE/UFRJ. Rio de Janeiro. 2013.

MATSUSHITA, R. Y. – **O que é Estatística?** Disponível em: <http://vsites.unb.br/ie/est/complementar/estatistica.htm>

MCKINSEY, ***'Game changers: Five opportunities for US growth and renewal'***, 2013. Disponível em: http://www.mckinsey.com/~/media/McKinsey/Global%20Themes/Americas/US%20game%20changers/MGI_Game_changers_US_growth_and_renewal_Full_report.ashx

MCKINSEY, ***'Big Data: The next frontier for innovation, competition and productivity'***, 2011. Disponível em: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

MENDES, Daniel**. Fundamentos de *Big Data***, 2015. Disponível em: https://www.datascienceacademy.com.br/course?courseid=big-data-fundamentos

MODESTO, L. R. **Representação e Persistência para acesso a Recursos Informacionais Digitais gerados dinamicamente em sítios oficiais do Governo Federal**. 2013. 103 f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília. 2013.

MORETTIN, P. A. – **Introdução à estatística para ciências exatas**. São Paulo: Atual, 1981.

O'BRIAN, J. A. – **Sistemas de Informação e as Decisões Gerenciais na era da Internet**. 2ª ed., São Paulo: Saraiva, 2004.

RIBEIRO, C. J. S. – ***Big Data*: os novos Desafios para os Profissionais da Informação**. Rio de Janeiro. UNIRIO. 2014

RIBEIRO, C. J. S. – **Diretrizes para o Projeto de Portais de Informação: Uma Proposta Interdisciplinar Baseada na Análise de Domínio e Arquitetura da Informação**. 2008. Disponível em: http://enancib.ibict.br/index.php/enancib/xenancib/paper/viewFile/3369/2495

SALSBURG, D. – **Como a Estatística Revolucionou a Ciência no Século XX**. Rio de Janeiro: Zahar, 2009.

SARTORIS, A. – **Estatística e Introdução à Econometria.** São Paulo. Saraiva, 2003.

SETZER, V. - **Dado, Informação, Conhecimento e Competência.** 1999. Disponível em: https://www.ime.usp.br/~vwsetzer/datagrama.html

SEYMOUR, C. – ***The State of Big Data*.** Disponível em: http://www.econtentmag.com/Articles/Editorial/Feature/The-State-of-Big-Data-2017-115703.htm

SILVEIRA, R. D. F. – **Mineração de Dados Aplicada à Definição de Índices em Sistemas de Raciocínio Baseado em Caos**. UFRGS, 2003.

STIGLER, S. M. – ***The history of Statistics: The Measurement of Uncertainty Before* 1900**. Cambridge: Belknap Press of Harvard University Press, 1986.

TAURION, Cezar. - **Volume, variedade, velocidade, veracidade e valor: Os cinco V's do *Big Data***, 2016. Disponível em: http://computerworld.com.br/volume-variedade-velocidade-veracidade-e-valor-os-cinco-vs-do-big-data?

TENER, O - ***Privacy and Big Data: The Biggest Public Policy Challenge of Our Time?***, 2013. Disponível em: https://iapp.org/news/a/privacy-and-big-data-making-ends-meet/

VARIAN, HAL R. ***Big Data: New Tricks for Econometrics.*** *Journal of Economic Perspectives,* Vol. 28, 2014.

WILLS, J. ***7 Ways Amazon Uses Big Data to Stalk You (AMZN).*** Investopedia, 2016. Disponível em: http://www.investopedia.com/articles/insights/090716/7-ways-amazon-uses-big-data-stalk-you-amzn.asp